RICE UNIVERSITY

An Approach for the Adaptive Solution of Optimization Problems Governed by Partial Differential Equations with Uncertain Coefficients

by

Drew P. Kouri

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Thesis Committee:

Matthias Heinkenschloss, Chairman
Professor of Computational and Applied Mathematics

Danny C. Sorensen
Noah G. Harding Professor of Computational and Applied Mathematics

Beatrice M. Riviere
Associate Professor of Computational and Applied Mathematics

Dennis D. Cox
Professor of Statistics

Houston, Texas

May, 2012

Abstract

An Approach for the Adaptive Solution of Optimization Problems Governed by Partial Differential Equations with Uncertain Coefficients

by

Drew P. Kouri

In this thesis, I develop and analyze a general theoretical framework for optimization problems governed by partial differential equations (PDEs) with random inputs. This theoretical framework is based on the adjoint calculus for computing derivatives of the objective function. I develop an efficient discretization and numerical optimization algorithm for the solution of these PDE constrained optimization problems. Using derivative based numerical optimization algorithms to solve these PDE constrained optimization problems is computationally expensive due to the large number of PDE solves required at each iteration. I present a stochastic collocation discretization for these PDE constrained optimization problems and prove the convergence of this discretization method for a specific class of problems.

The stochastic collocation discretization technique described here requires many decoupled PDE solves to compute gradient and Hessian information. I develop a novel optimization theoretic framework based on dimension adaptive sparse grid quadrature to reduce the total number of PDE solves. My adaptive framework employs basic or retrospective trust regions to manage the adapted stochastic collocation models. In addition, I prove global first order convergence of the retrospective trust region

algorithm under weakened assumptions on the modeled gradients. In fact, if one can bound the error between actual and modeled gradients using reliable and efficient *a posteriori* error estimators, then the global convergence of the retrospective trust region algorithm follows.

Finally, I describe a high performance implementation of my adaptive collocation and trust region framework. This framework can be efficiently implemented in the C++ programming language using the Message Passing Interface (MPI). Due to the large number of PDE solves required for derivative computations, it is essential to choose inexpensive approximate models and appropriate large-scale nonlinear programming techniques throughout the optimization routine to obtain an efficient algorithm. Numerical results for the adaptive solution of these optimization problems are presented.

Acknowledgements

I would like to give my sincere gratitude to my adviser, Matthias Heinkenschloss, for giving me the opportunity to work on this problem. His assistance and guidance over the past four years has opened many doors for me. Furthermore, I am grateful for my doctoral committee: Danny Sorensen, Beatrice Riviere, and Dennis Cox. Their careful reading of this thesis and many great suggestions have been critical to the success of this thesis.

I would like to thank Denis Ridzal for giving me the opportunity to work at Sandia National Laboratory. His assistance in my high performance implementation of the methods described within this thesis and his perspective on PDE constrained optimization has been crucial. In addition, I would like to thank Bart van Bloemen Waanders at Sandia National Laboratory for the many helpful discussions on mathematics as well as my future postdoctoral opportunities.

Above all, this thesis is dedicated to my family. Their constant love and encouragement has aided in all of my accomplishments.

This research was supported by AFOSR grant FA9550-09-1-0225 and NSF grant DMS-0915238.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advances in computational power and efficient numerical simulation of complex physical systems, simulation based optimization and uncertainty quantification are becoming increasingly feasible. The ability to simulate complex behaviors of physical systems gives engineers and experimental scientists the ability to make predictions and hypotheses about certain outputs of interest. These simulations are critical for scientists of many fields such as: fluid dynamics, heat transfer, chemically reacting systems, oil field research, nonproliferation seismology, monitoring of CO_2 output, radiation transport, climate science, and structural mechanics. With these computational and mathematical forward models come uncertainty concerning computed quantities [83, 84]. In physical modeling and numerical simulation, uncertainty arises from lack of knowledge corresponding to model physics and assumptions (epistemic uncertainty) and inherent variability in the model parameters (aleatory uncertainty). Epistemic uncertainty can be reduced by increased knowledge and is generally not probabilistic in nature, although one can, to some extent, handle epistemic uncertainty in the Bayesian framework. On the other hand, aleatory uncertainty is typically characterized by assigning probability distributions to uncertain or random parameters. The quantification of aleatory uncertainty is performed by tracking these probability distributions through the forward simulations.

This work is focused on the treatment of aleatory uncertainty. Common techniques for quantifying aleatory uncertainty are intrusive and non-intrusive expansion methods [46]. Intrusive methods include local and global polynomial chaos expansion methods [42, 10, 70]. Such methods seek a global or local polynomial representation of the outputs of interest. The name intrusive refers to the fact that these methods are typically not "black box;" that is, numerical implementation of such methods requires changes to black box forward simulation code [83]. Moreover, intrusive methods typically result in large coupled systems to be solved. Common non-intrusive methods are sample based collocation methods such as (quasi-)Monte Carlo and stochastic collocation [121, 123, 9]. Non-intrusive methods result in many decoupled deterministic forward simulations at random or structured collocation points.

The stochastic collocation method seeks an interpolated polynomial representation of the random field output of interest. Furthermore, provided sufficient regularity of the output of interest, stochastic collocation enjoys rapid convergence as a discretization scheme. This is contrary to Monte Carlo methods. Monte Carlo methods converge in probability at a rate of $1/\sqrt{Q}$ where Q is the number of random samples. This means, one requires a large number of samples to decrease the error in simulation. Aside from the slow convergence rate, Monte Carlo avoids the curse of dimensionality, i.e. the convergence rate is independent of the dimension of the problem. Stochastic collocation on the other hand exploits regularity of the output of interest to speed convergence [9]. Unlike Monte Carlo methods, stochastic collocation does not thwart the curse of dimensionality, although the dependence of the convergence rate on the stochastic dimension can be reduced by employing sparse grids [52, 87, 88, 15, 119, 32, 89, 92, 93]. Sparse grids were first introduced by Smolyak in 1963 [110] and present an efficient means of approximating tensor product problems. Sparse grids and stochastic collocation converge rapidly for sufficiently smooth problems and are vastly superior to Monte Carlo methods for problems with a modest number of random variables.

Forward simulation and uncertainty quantification are typically not the only goals of experimentalists. In many applications, these forward simulation codes are used to aid in control and design of physical processes. Also, these simulators can be used to infer parameters of the physical system. The discretization of these problems results in extremely large scale constrained optimization problems even when uncertainty is not considered. When uncertainty is added to the forward problems, the scale of the resulting optimization problem increases drastically. Moreover, the addition of uncertainty requires problem reformulation. These reformulations result in "robust" optimization problems where the goal is to minimize the risk associated with a certain design or control. Here, risk refers to a measure of the variation of the objective with respect to the random model input data. These measures of risk are studied in depth in the context of stochastic programming. A particularly convenient and rich class of risk measures are the coherent risk measures [98, 99, 108]. These measures exhibit properties desirable for optimization, such as convexity and monotonicity. Reformulating the optimization problem may also result in chance or probabilistic constraints where one requires that the probability that certain outputs of interest exceed set thresholds be less than a set "tolerable" level [116]. These constraints may arise as constraints on the probability of failure of the optimal design.

Risk measures and chance constraints add much complexity to the analytical and computational aspects of these optimization problems. Naïvely applied, these additions typically result in a lack of differentiability for the objective function and constraints [108]. To this end, traditional gradient based optimization methods may not apply. Due to the extreme computational complexity of these problems, it is advantageous to use gradient based methods to generate reliable and efficient optimization routines. In order to apply gradient based methods, one may need to choose a suitable smooth alternate to the risk measure or chance constraints, or one may need to reformulate the optimization problem by adding auxiliary variables. Care must be taken when reformulating these problems and discretizing using stochastic collocation

on sparse grids, which produce negative quadrature weights, as these reformulations may lead to inconsistencies, i.e. loss of convexity of the objective or constraints. The current treatment of risk measures and chance constraints typically involves the use of Monte Carlo methods and to my knowledge has not been tackled using polynomial chaos or stochastic collocation methods.

Aside from problem formulation issues, optimization problems governed by uncertain forward models require possibly many simulations of the forward model per gradient based optimization iteration. Furthermore, gradient based methods require some control over the forward simulation code in order to compute adjoint states and gradients. Finite differences can be employed to avoid modifying black box forward simulation code, but finite difference computations become prohibitive for large scale optimization problems as each gradient computation requires multiple forward simulations. I consider adjoint based optimization because most models used in engineering lend themselves to adjoint computations. The adjoint approach yields an efficient and accurate gradient computation at the expense of destroying the black box quality of the forward simulator. When non-intrusive uncertainty quantification methods are used with the adjoint approach, the forward and adjoint problems at each sample can be solved in parallel using the deterministic solvers, i.e. no new code needs to be generated to accommodate uncertainty when computing gradients [67].

Even the use of adjoint based gradient computation does not solve the problem of algorithmic and computational inefficiency. As stated above, adjoint based computations require many forward simulations with non-intrusive methods or, in the case of intrusive methods, the solution of many large coupled systems of equations. Hence, optimization quickly becomes prohibitive. To circumvent this, I propose to use adaptive uncertainty quantification methods to solving these optimization problems under uncertainty. Optimization is an ideal setting for adaptivity as accuracy in model simulation is only essential when close to the minimizer. Moreover, optimization gives a natural metric to guide adaptivity. I have developed an adaptive sparse grid stochastic

collocation framework to solve these optimization problems. My adaptive approach is based on the trust region algorithm for unconstrained optimization, which provides an exemplary optimization theoretic framework for managing approximate models. In this case, the adaptivity is driven according to the size of the gradient. Therefore, as one approaches a first order critical point, the model is increasingly refined. Furthermore, it is possible to extend the trust region idea to some classes of constrained problems in which case the adaptivity is driven by the projected gradient. To this end, I propose to extend my adaptive sparse grid collocation trust region framework to handle more general constrained and chance constrained optimization problems. The goal of this extension is to develop efficient algorithms and software to handle challenging engineering application problems.

1.1 Literature Review

My work presented in this thesis lies in the intersection of many mathematical fields such as optimization theory, probability theory, approximation theory, and the theory of partial differential equations (PDEs). It is my goal in this section to review some relevant background material concerning PDEs with uncertain coefficients, sparse grids, optimization problems governed by PDEs with uncertain coefficients, trust region methods, and adaptive finite element methods.

1.1.1 PDEs with Uncertain Coefficients

The study of PDEs with uncertain coefficients is a relatively new subject, but the building blocks for the modern analysis of such PDEs were formed in the early twentieth century with Norbert Wiener's 1938 development of "homogeneous chaos" or polynomial chaos expansion [122]. This polynomial chaos expansion provides a polynomial representation of Gaussian random fields and remains a popular numerical method of solving PDEs with uncertain coefficients [65, 72, 113]. Other contributions

to the modern theory of PDEs with uncertain coefficients stems from the independent works of Karhunen [64] and Lòeve [71] who developed a general Fourier series representation of random fields known today as the Karhunen-Lòeve (KL) expansion. More recently, polynomial chaos and the KL expansion have been employed to create numerical methods for solving PDEs with uncertain coefficients.

The polynomial chaos methods for solving PDEs with uncertain coefficients has received much attention [54, 124, 65]. Coupling the polynomial chaos method with finite elements or other numerical PDE solution techniques gives a robust numerical solution method, but suffers from the need to compute with high order global polynomials. Babuska, Tempone, and Zouraris [11] developed a finite element scheme known as stochastic Galerkin which encompasses polynomial chaos and allows for local (discontinuous) polynomial representations of the random field solution. The stochastic Galerkin method is very popular due to its rapid convergence rates, but suffers from high computational costs. The stochastic Galerkin discretization results in a large coupled linear system which may be expensive to solve.

Another class of methods known as sample based or intrusive methods circumvent the need to solve a large coupled linear system. This class of methods contains Monte Carlo, quasi Monte Carlo, and stochastic collocation. These methods lead to many decoupled linear systems which may be solved in parallel. The Monte Carlo and quasi Monte Carlo methods have their roots in probability theory, whereas the collocation method has its roots in approximation theory. The stochastic collocation method seeks to interpolate the random field PDE solution on a set of quadrature nodes [123, 121, 86, 8]. Aside from being possibly more efficient than the stochastic Galerkin method, the stochastic collocation method also enjoys similar convergence properties to stochastic Galerkin [14].

1.1.2 Sparse Grids and Approximation Theory

As mentioned, stochastic collocation discretization results in a decoupled system of PDEs. The number of PDEs to be solved is exactly the number of nodes used for the interpolation of the PDE solution. Therefore, it is essential to choose sets of nodes which are small in size and exhibit high polynomial accuracy. These requirements warrant the use of so called sparse grids. The sparse grid idea was developed in 1963 by Smolyak [111]. Recently, sparse grids have grown in popularity. As such, the convergence of sparse quadrature and interpolation rules is well known [52, 87, 88, 15, 119, 32, 89, 92, 93]. General sparse grid rules achieve similar orders of accuracy as corresponding tensor product rules, but have far fewer nodes. In fact, if the sparse grid is based on nested one dimensional rules, then the sparse grid is even sparser (i.e. has even fewer nodes). Some common choices of nested rules are the Clenshaw-Curtis rule for uniformly distributed random variables [40], or any Gauss-Patterson rule [91, 51]. Recently, many researchers have investigated adaptive and optimized sparse grids [53, 55, 56], where the goal is to choose a sparse grid rule which is both optimal in accuracy and number of nodes given a certain class of functions.

1.1.3 PDE Constrained Optimization Under Uncertainty

The subject of optimization problems governed by PDEs with uncertain coefficients lies at the interface of PDEs with uncertain coefficients, optimization in Banach spaces, and stochastic programming. Stochastic programming offers many numerical schemes for solving problems with uncertainty, such as the sample average approximation (SAA) and the stochastic approximation algorithm (SA) [90, 108, 75]. These methods are Monte Carlo based methods and thus, not central to this thesis. Stochastic programming also gives the framework for dealing with probabilistic (or chance) constraints, as well as, develops the theory of coherent risk measures [100, 98, 116, 115, 114]. Both probabilistic constraints and coherent risk measures are vastly important to the topics of this thesis.

Combining ideas from stochastic programming and PDE constrained optimization [63] gives the rich theory of PDE optimization under uncertainty. Although there has been much work in the field of statistical inverse problems and signal recovery [74, 76, 118, 22], very few researchers actually consider optimal control of uncertain PDEs. The few works that have considered such control problems lack a complete theoretical framework and in some situations, do not seek necessarily optimal solutions [104, 25, 105, 27, 24, 26]. One common approach in these sources solves the control problem as a sequence of deterministic problems defined at the collocation points. The controls are then taken as the expected value of the computed controls. Controls generated in this way are not necessarily optimal. Thus, there is a strong need for a unified understanding and theory of PDE optimization under uncertainty. Some of this theory can be found in [67].

1.1.4 Trust Regions and Frameworks for Adaptivity

As hinted at earlier, there is great need for efficient and reliable numerical methods for the solution of optimization problems governed by PDEs with uncertain coefficients. The goal of such methods should be to solve optimization problems using inexpensive approximate models whenever possible. This framework is known as model management and is an inexpensive and efficient solution method for optimization problems governed by PDEs. Typically, model management frameworks are based on trust regions due to their flexibility and provable global convergence [2, 3, 43, 44]. In fact, one can prove convergence of trust region methods with minimal conditions on function and gradient exactness [36, 81, 96, 41, 60]. Such flexibility is ideal for model adaptation. This idea of model adaptive trust regions has recently been exploited in the context SQP methods in [127]. Furthermore, in [16], this idea has been applied to the solution of stochastic programs using Monte Carlo sampling techniques. In this thesis, I consider a novel trust region approach known as the retrospective trust region method [17]. The retrospective trust region updates the trust region radius

following model updates (as opposed to before). As such, the trust region radius is updated to the new model rather than the old model. This modification may decrease the possibility of prohibitively small trust region radii.

The quality of these adaptive frameworks is contingent on the quality of error estimators used for adaptation. In the case of optimization problems governed by PDEs with uncertain coefficients, there are three possible sources of error to be controlled. First of all, the spatial PDE discretization can be adaptively controlled using adaptive finite elements based on well known *a posteriori* error indicators [33, 34, 95, 18, 19, 20, 61, 62]. Another source of error in the stochastic collocation finite element solution arises from the quadrature rule used in defining the collocation space. Few attempts have been made to control this error adaptively [53, 73, 1] and it would be desirable to improve these estimates. A final possible source of error is in model order reduction in the case of time dependent PDE constraints [21, 57, 4]. Although model reduction is not the main topic of this thesis, the trust region framework presented here is valid for adaptive model reduction in PDE constrained optimization.

1.2 Thesis Outline

In this thesis, I will first formulate a general form for optimization problems governed by PDEs with uncertain coefficients. In this formulation, I will give assumptions on the objective function and PDE constraint that guarantee well-posedness of the optimization problem. Furthermore, I will introduce a model problem for this thesis which corresponds to the distributed control of an elliptic PDE. Next, I will develop the stochastic collocation method for solving PDEs with uncertain coefficients and extend the collocation method to the case of optimization. Additionally I will prove an *a priori* error bound for a class of control problems solved using the sparse grid stochastic collocation finite element method. Following this collocation discussion, I will develop general sparse grids for high dimensional interpolation and quadrature.

In my discussion of sparse grids, I prove new interpolation and approximation results for general sparse grid operators. Subsequently, I will present a novel model adaptive trust region approach to solving optimization problems governed by PDEs with uncertain coefficients. Here, I analyze two trust region frameworks: the basic trust region algorithm and the retrospective trust region algorithm. For the basic trust region algorithm, my model adaptive framework is provably global convergent due to results found in [60]. On the other hand, I prove the global convergence of the retrospective trust algorithm under a weakened gradient conditions. Finally, I will provide a brief discussion of implementation details and present numerical results.

Chapter 2

Optimization Under Uncertainty

Uncertainty is present in nearly every physical system and in many engineering applications the risk-averse optimization of such systems is crucial. In the literature, the concept of risk is predominately applied to the optimization of financial portfolios. My goal is to extended the concept of risk-averse optimization to engineering and science applications. For example, in engineering design and control problems, risk-averse optimization can be used as a certificate of reliability. In this chapter, I will present a general formulation of the risk-averse optimization problem. I will develop the test problem that I will consider throughout this thesis. In the deterministic setting, this test problem is a quadratic program posed in a Banach space. An instance of this test problem is the quadratic optimal control of linear elliptic partial differential equations (PDEs) with uncertain coefficients. This test problem is very insightful and sheds light on a general formulation for these optimization problems. Following the test problem I will formulate the general problem setting in which I will study these risk-averse optimization problems and develop my adaptive stochastic collocation and trust region framework. Moreover, I will formulate assumptions on the general optimization problem. These assumptions will be used to ensure existence and uniqueness of optimal solutions as well as ensure that the stochastic collocation method is an applicable discretization technique. I will then derive the adjoint calcu-

lus for computing derivatives of the risk-averse objective function and present some standard results concerning tensor products of Banach spaces.

2.1 Test Problem

Let $\mathcal{H}$ and $\mathcal{Z}$ be Hilbert spaces, and let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces. I will begin this chapter by presenting the archetypal test problem used throughout this thesis. This optimization problem is an extension of the deterministic quadratic program

$$\begin{array}{lrcl} \min\limits_{v\in\mathcal{V},\, z\in\mathcal{Z}} & j(v,z) \;:=\; \dfrac{1}{2}\|\mathbf{Q}v-\bar{q}\|_{\mathcal{H}}^2 & + & \dfrac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 \\ \text{subject to} & \mathbf{A}v+\mathbf{B}z+\mathbf{b} & = & 0, \end{array} \tag{2.1.1}$$

where $\mathbf{Q} \in \mathcal{L}(\mathcal{V},\mathcal{H})$ is an observation operator, $\bar{q} \in \mathcal{H}$ is the desired state of the system, $\mathbf{A} \in \mathcal{L}(\mathcal{V},\mathcal{W})$ is the state operator, $\mathbf{B} \in \mathcal{L}(\mathcal{Z},\mathcal{W})$ is the control operator, and $\mathbf{b} \in \mathcal{W}$ is an inhomogeneity.

2.1.1 Optimal Control of PDEs

An instance of (2.1.1) of particular interest to this thesis is the optimal control of the linear elliptic PDE

$$\begin{aligned} -\nabla\cdot(\epsilon(x)\nabla v(x)) &= z(x), \quad x\in D \\ v(x) &= 0, \qquad x\in\partial D, \end{aligned} \tag{2.1.2}$$

where $D \subset \mathbb{R}^d$ for $d = 1,\, 2,\, 3$ denotes the physical domain. If v is a solution to (2.1.2) and w is a sufficiently regular test function such that $w(x) = 0$ for all $x \in \partial D$, then integration by parts results in the variational problem:

Given $z \in \mathcal{Z} \subseteq H^{-1}(D)$, find $v \in \mathcal{V} := H_0^1(D)$ such that

$$\int_D \epsilon(x)\nabla v(x)\cdot\nabla w(x)\mathrm{d}x = \int_D z(x)w(x)\mathrm{d}x \quad \forall\, w\in\mathcal{V}. \tag{2.1.3}$$

For simplicity, let $\mathcal{Z} = L^2(D)$. Defining $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and $b : \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$ as

$$a(v, w) := \int_D \epsilon(x) \nabla v(x) \cdot \nabla w(x) \mathrm{d}x \quad \text{and} \quad b(w; z) := \int_D z(x) w(x) \mathrm{d}x, \tag{2.1.4}$$

the variational problem, (2.1.3), can be written in the equivalent form:

$$\text{Given } z \in \mathcal{Z}, \text{ find } v \in \mathcal{V} \text{ such that} \quad a(v, w) = b(w; z) \quad \forall\, w \in \mathcal{V}.$$

Furthermore, assuming $\epsilon \in L^\infty(D)$ is bounded away from zero almost everywhere in D, i.e.

$$\exists\, \epsilon_{\min} > 0 \quad \text{such that} \quad \epsilon(x) \geq \epsilon_{\min} \quad \text{a.e. in } D,$$

then a is a coercive and continuous bilinear form and hence the Lax-Milgram Theorem (Theorem 2.7.7 in [29]) ensures the existence of a unique $v(z) = v \in \mathcal{V}$ which solves (2.1.3).

Now, since $a(v, \cdot) \in \mathcal{V}^*$ for all $v \in \mathcal{V}$ and the mapping $v \in \mathcal{V} \mapsto a(v, \cdot) \in \mathcal{V}^*$ is continuous and linear, there exists a bounded linear operator $\mathbf{A} \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ such that

$$a(v, w) = \langle \mathbf{A}v, w \rangle_{\mathcal{V}^*, \mathcal{V}} \quad \forall\, v, w \in \mathcal{V},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{V}^*, \mathcal{V}}$ denotes the duality pairing on $\mathcal{V}$. For a more thorough discourse on the existence of $\mathbf{A}$ see Chapter 1.3 in [63]. Similar arguments prove the existence of the bounded linear operator $\mathbf{B} \in \mathcal{L}(\mathcal{Z}, \mathcal{V}^*)$ satisfying

$$b(w; z) = -\langle \mathbf{B}z, w \rangle_{\mathcal{V}^*, \mathcal{V}} \quad \forall\, z \in \mathcal{Z},\ w \in \mathcal{V}.$$

Therefore, the PDE (2.1.3) has the same form as the linear constraint in (2.1.1). The Lax-Milgram Theorem ensures the invertibility of the operator, $\mathbf{A}$, and thus the solution to (2.1.3) has the specific form

$$v(z) = -\mathbf{A}^{-1}\mathbf{B}z = \mathbf{S}z$$

where $\mathbf{S} \in \mathcal{L}(\mathcal{Z}, \mathcal{V})$ denotes the solution operator. Note that the solution $v(z) = v \in \mathcal{V}$ of (2.1.3) depends linearly on the control variable, $z \in \mathcal{Z}$.

Now for any real Hilbert space, $\mathcal{H} \supseteq \mathcal{V}$, the optimal control problem is

$$\begin{aligned} \min_{v\in\mathcal{V},\, z\in\mathcal{Z}} \qquad & j(v,z) \;:=\; \frac{1}{2}\|v-\bar{v}\|_{\mathcal{H}}^2 \;+\; \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 \\ \text{subject to} \qquad & \mathbf{A}v+\mathbf{B}z \;=\; 0. \end{aligned} \tag{2.1.5}$$

Relating (2.1.5) to (2.1.1), $\mathcal{W} = \mathcal{V}^*$, $\mathbf{Q}$ is the identity operator, and $\bar{v} = \bar{q}$. To make this concrete, I will set $\mathcal{H} = L^2(D)$. Furthermore, the assumptions on ϵ permit the reformulation of (2.1.5) as the implicitly constrained optimization problem [59]

$$\min_{z\in\mathcal{Z}} \widehat{j}(z) := j(v(z),z) = \frac{1}{2}\|\mathbf{S}z-\bar{v}\|_{\mathcal{H}}^2 + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2. \tag{2.1.6}$$

2.1.2 PDEs with Random Inputs

Similar analysis as above holds when ϵ in (2.1.3) is replaced by a random field (i.e. a family of coefficients, ϵ, indexed by a random variable). Let $(\Omega, \mathcal{F}, P)$ denote a complete probability space. Ω is the set of outcomes, $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra of events, and $P : \mathcal{F} \to [0,1]$ is a probability measure. The random field coefficient is defined as $\epsilon : \Omega \times D \to \mathbb{R}$ where the map $\omega \in \Omega \mapsto \epsilon(\omega, \cdot)$ is P-measurable. Furthermore, to extend the existence and uniqueness result from the deterministic case, it suffices to assume that $\epsilon \in L^{\infty}(\Omega \times D)$ is bounded away from zero almost everywhere in $\Omega \times D$, i.e.

$$\exists\, \epsilon_{\min} > 0 \quad \text{such that} \quad \epsilon(\omega,x) \geq \epsilon_{\min} \quad \text{a.e. in } \Omega \times D. \tag{2.1.7}$$

For weaker assumptions on ϵ see Section 1 of [9]. Substituting this random field coefficient into (2.1.3) gives rise to the PDE

$$\begin{aligned} -\nabla\cdot(\epsilon(\omega,x)\nabla u(\omega,x)) = z(x) \quad & x \in D, \text{ a.e. in } \Omega \\ u(\omega,x) = 0 \qquad & x \in \partial D, \text{ a.e. in } \Omega. \end{aligned} \tag{2.1.8}$$

As indicated by the notation, $u(\omega, x)$, the solution to the state equation, (2.1.8), is also a random field. Moreover, for almost all $\omega \in \Omega$, the solution to (2.1.8) satisfies $u(\omega) \in \mathcal{V}$.

In order to discuss the weak form of (2.1.8), I will require the notion of a Bochner space. Bochner spaces are formal extensions of the Lebesgue spaces, $L^p_P(\Omega)$, for functions which output into general Banach spaces [125]. The Bochner space $L^p_P(\Omega;\mathcal{V})$ is defined as

$$L^p_P(\Omega;\mathcal{V}) := \left\{ v : \Omega \to \mathcal{V} \ : \ v \text{ strongly measurable}, \ \int_\Omega \|v(\omega)\|^q_{\mathcal{V}} \mathrm{d}P(\omega) < +\infty \right\}$$

for $p \in [1,\infty)$ and

$$L^\infty_P(\Omega;\mathcal{V}) := \{ v : \Omega \to \mathcal{V} \ : \ v \text{ strongly measurable}, \ \text{ess-sup}_{\omega\in\Omega} \|v(\omega)\|_{\mathcal{V}} < +\infty \}$$

for $p = \infty$. Returning to (2.1.8), assumption (2.1.7) and the Lax-Milgram Theorem ensure the existence of a unique solution, $u \in L^2_P(\Omega;\mathcal{V})$, which solves the variational problem:

Given $z \in \mathcal{Z}$, find $u \in L^2_P(\Omega;\mathcal{V})$ such that

$$\int_\Omega \int_D \epsilon(\omega,x) \nabla u(\omega,x) \cdot \nabla w(\omega,x) \mathrm{d}x \mathrm{d}P(\omega) = \int_\Omega \int_D z(x) w(\omega,x) \mathrm{d}x \mathrm{d}P(\omega) \quad \forall w \in L^2_P(\Omega;\mathcal{V}). \tag{2.1.9}$$

2.1.2.1 The Finite Noise Assumption

To facilitate the numerical solution of (2.1.3), I will work under the finite noise assumption. Assume there exists a finite vector of M independent random variables $Y : \Omega \to \Gamma$ with $\Gamma := \Gamma_1 \times \ldots \times \Gamma_M$ with $\Gamma_k \subset \mathbb{R}$ for $k = 1,\ldots,M$ such that $\epsilon(\omega,\cdot) \equiv \epsilon(Y(\omega),\cdot)$. Furthermore, suppose that each random variable Y_k for $k = 1,\ldots,M$ has Lebesgue density $\rho_k : \Gamma_k \to [0,\infty]$ and the vector Y has joint density $\rho(y) = \rho_1(y_1) \cdot \ldots \cdot \rho_M(y_M)$. In this case, the probability measure $dP(\omega)$ can be replaced by the weighted Lebesgue measure $\rho(y)\mathrm{d}y$. Additionally, this assumption

permits the change of variables in (2.1.9):

$$\text{Given } z \in \mathcal{Z}, \text{ find } u \in L^2_\rho(\Gamma;\mathcal{V}) \text{such that}$$
$$\int_\Gamma \rho(y) \int_D \epsilon(y,x)\nabla u(y,x)\cdot\nabla w(y,x)\mathrm{d}x\mathrm{d}y = \int_\Gamma \rho(y)\int_D z(x)w(y,x)\mathrm{d}x\mathrm{d}y \quad \forall w \in L^2_\rho(\Gamma;\mathcal{V}) \tag{2.1.10}$$

where the ρ-weighted Bochner spaces are defined analogously as

$$L^p_\rho(\Gamma;\mathcal{V}) := \left\{ v : \Gamma \to \mathcal{V} \ : \ v \text{ strongly measurable, } \int_\Gamma \rho(y)\|v(y)\|^q_{\mathcal{V}}\mathrm{d}y < +\infty \right\}$$

for $p \in [1,\infty)$ and

$$L^\infty_\rho(\Gamma;\mathcal{V}) := \left\{ v : \Gamma \to \mathcal{V} \ : \ v \text{ strongly measurable, } \text{ess-sup}_{y\in\Gamma}\, \rho(y)\|v(y)\|_{\mathcal{V}} < +\infty \right\}.$$

for $p = \infty$. If $\mathcal{V}^*$ is separable, then for $p \in (1,\infty)$, the topological dual spaces corresponding to $L^p_\rho(\Gamma;\mathcal{V})$ is isometrically isomorphic with $L^{p^*}_\rho(\Gamma;\mathcal{V}^*)$ where $\frac{1}{p}+\frac{1}{p^*} = 1$. For $p = 1$, the topological dual of $L^1_\rho(\Gamma;\mathcal{V})$ is isometrically isomorphic to $L^\infty_\rho(\Gamma;\mathcal{V}^*)$, but the same is not true for $p = \infty$. In this case, $L^\infty_\rho(\Gamma;\mathcal{V})^* \supset L^1_\rho(\Gamma;\mathcal{V}^*)$ [37, 125, 126].

The uniform ellipticity (2.1.7) for (2.1.3) can be restated in the image space Γ as

$$\exists\ \epsilon_{\min} > 0 \quad \text{such that} \quad \epsilon(y,x) \geq \epsilon_{\min} \quad \text{a.e. in } \Gamma \times D. \tag{2.1.11}$$

This condition ensures existence and uniqueness of solutions to the variational problem (2.1.10), but also implies additional regularity of the solution with respect to $y \in \Gamma$. Assumption 2.1.11 can be used to prove the continuity result

$$\|u(y)\|_{H^1(D)} \leq C\|z\|_{\mathcal{Z}} \quad \text{a.e. in } \Gamma$$

where C is some positive constant independent of $y \in \Gamma$ (see [9]). Since the right hand side is independent of $y \in \Gamma$, this inequality implies $u \in L^\infty_\rho(\Gamma;\mathcal{V})$.

2.1.2.2 The Weak Form

Define the bilinear form $a : \mathcal{V}\times\mathcal{V}\times\Gamma \to \mathbb{R}$ as

$$a(v,w;y) = \int_D \epsilon(y,x)\nabla v(x)\cdot\nabla w(x)\mathrm{d}x,$$

the linear form $b : \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$ as in (2.1.4), and the expected value operator, $E : L^1_\rho(\Gamma) \to \mathbb{R}$, as

$$E[X] = \int_\Gamma \rho(y) X(y) \mathrm{d}y.$$

Then, the weak form, (2.1.9), can be equivalently written as:

$$\begin{aligned} &\text{Given } z \in \mathcal{Z}, \text{ find } u \in L^2_\rho(\Gamma; \mathcal{V}) \text{ such that} \\ &E[a(u, w; \cdot)] = E[b(w; z)] \quad \forall\, w \in L^2_\rho(\Gamma; \mathcal{V}). \end{aligned} \tag{2.1.12}$$

Throughout this thesis the weak form of the state equation will be denoted $e : L^2_\rho(\Gamma; \mathcal{V}) \times \mathcal{Z} \to L^2_\rho(\Gamma; \mathcal{V})^*$ which, in this case, is defined as

$$e(u, z) := \int_\Gamma \rho(y) \Big\{ a(u(y), \cdot\,; y) - b(\,\cdot\,; z) \Big\} \mathrm{d}y.$$

The uniform ellipticity assumption (2.1.11) and the Lax-Milgram theorem ensures the existence of a unique solution to the variational problem (2.1.12), or equivalently

$$\begin{aligned} &\text{Given } z \in \mathcal{Z}, \text{ find } w \in L^2_\rho(\Gamma; \mathcal{V}) \text{ such that} \\ &\langle e(w, z), v \rangle_{L^2_\rho(\Gamma;\mathcal{V})^*, L^2_\rho(\Gamma;\mathcal{V})} = 0 \quad \forall v \in L^2_\rho(\Gamma; \mathcal{V}). \end{aligned} \tag{2.1.13}$$

Similar to the discussion concerning the deterministic problem, there exist bounded linear operators $\mathbf{A} \in \mathcal{L}(L^2_\rho(\Gamma; \mathcal{V}), L^2_\rho(\Gamma; \mathcal{V})^*)$ and $\mathbf{B} \in \mathcal{L}(\mathcal{Z}, L^2_\rho(\Gamma; \mathcal{V})^*)$ defined by

$$\langle \mathbf{A}u, w \rangle_{L^2_\rho(\Gamma;\mathcal{V})^*, L^2_\rho(\Gamma;\mathcal{V})} = \int_\Gamma \rho(y) a(u(y), w(y); y) \mathrm{d}y \quad \forall\, u, w \in L^2_\rho(\Gamma; \mathcal{V}) \tag{2.1.14a}$$

and

$$\langle \mathbf{B}z, w \rangle_{L^2_\rho(\Gamma;\mathcal{V})^*, L^2_\rho(\Gamma;\mathcal{V})} = -\int_\Gamma \rho(y) b(w(y); z) \mathrm{d}y \quad \forall\, w \in L^2_\rho(\Gamma; \mathcal{V}),\ z \in \mathcal{Z}, \tag{2.1.14b}$$

such that (2.1.13) has the form

$$\mathbf{A}u + \mathbf{B}z = 0,$$

The Lax-Milgram theorem ensures the invertibility of $\mathbf{A}$ and therefore the solution to (2.1.13) can be written as

$$u(z) = -\mathbf{A}^{-1}\mathbf{B}z = \mathbf{S}z$$

where $\mathbf{S} \in \mathcal{L}(\mathcal{Z}, L^2_\rho(\Gamma; \mathcal{V}))$ denotes the solution operator. As above, the state, $u(z)$, depends linearly on the control variable, $z \in \mathcal{Z}$.

2.1.2.3 The Parametric Weak Form

For the numerical solution of (2.1.12) it can be favorable to consider the following the parametrized variational problem:

Given $z \in \mathcal{Z}$, find $u : \Gamma \to \mathcal{V}$ such that

$$a(u(y), w; y) = b(w; z) \text{ a.e. in } \Gamma, \quad \forall\, w \in \mathcal{V}. \tag{2.1.15}$$

To be concise, I will use the notation $\tilde{e} : \mathcal{V} \times \mathcal{Z} \times \Gamma \to \mathcal{V}^*$ where

$$\tilde{e}(u(y), z; y) = a(u(y), \cdot\,; y) - b(\cdot\,; z) = 0$$

to denote the state equation. Assumption (2.1.11) and the Lax-Milgram theorem ensure the existence of a point-wise almost everywhere defined function, $u : \Gamma \to \mathcal{V}$, which solves the variational problem (2.1.15), or equivalently

Given $z \in \mathcal{Z}$, find $u : \Gamma \to \mathcal{V}$ such that

$$\langle \tilde{e}(u(y), z; y), w \rangle_{\mathcal{V}^*, \mathcal{V}} = 0 \quad \text{a.e.} \quad \forall\, w \in \mathcal{V}. \tag{2.1.16}$$

The continuity bound on the solution of (2.1.16) implies $u \in L^\infty_\rho(\Gamma; \mathcal{V})$ and finiteness of the probability measure $\rho(y)\mathrm{d}y$ ensures that $u \in L^p_\rho(\Gamma; \mathcal{V})$ for any $p \in [1, \infty) \cup \{\infty\}$. In this test case, the solution u of (2.1.16) is also a solution of (2.1.13) and, by uniqueness, these solutions coincide.

Define $\widehat{\mathbf{A}}(y) \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ and $\widehat{\mathbf{B}}(y) \in \mathcal{L}(\mathcal{Z}, \mathcal{V}^*)$ for almost all $y \in \Gamma$ as

$$\langle \widehat{\mathbf{A}}(y) v, w \rangle_{\mathcal{V}^*, \mathcal{V}} = a(v, w; y) \quad \forall\, v, w \in \mathcal{V} \text{ a.e. in } \Gamma$$

and

$$\langle \widehat{\mathbf{B}}(y) z, w \rangle_{\mathcal{V}^*, \mathcal{V}} = -b(w; z) \quad \forall\, w \in \mathcal{V},\ z \in \mathcal{Z} \text{ a.e. in } \Gamma,$$

respectively. Then, the weak form, (2.1.15), gives rise to the linear operator equation

$$\widehat{\mathbf{A}}(y) u(y) + \widehat{\mathbf{B}}(y) z = 0 \quad \text{a.e. in } \Gamma.$$

The Lax-Milgram theorem ensures the point-wise almost everywhere invertibility of $\widehat{\mathbf{A}}$ and therefore the solution to (2.1.15) can be written as

$$u(y;z) = -\widehat{\mathbf{A}}^{-1}(y)\widehat{\mathbf{B}}(y)z = \widehat{\mathbf{S}}(y)z \quad \text{a.e. in } \Gamma$$

where $\widehat{\mathbf{S}}(y) \in \mathcal{L}(\mathcal{Z},\mathcal{V})$ for almost all $y \in \Gamma$ denotes the solution operator.

Now for $p \in [1,\infty)$ let

$$\mathcal{U} = L^p_\rho(\Gamma;\mathcal{V}) = L^p_\rho(\Gamma;H^1_0(D))$$

The parameter p of integrability will be chosen in the next section depending on the risk measure used in the objective function of the optimal control problem. One can generalize the operators in (2.1.14) as follows. Define

$$\mathcal{Y} = L^p_\rho(\Gamma;\mathcal{V}^*).$$

Note that $\mathcal{Y}^* = L^{p^*}_\rho(\Gamma;\mathcal{V})$ where $1/p + 1/p^* = 1$. Let $\mathbf{A} \in \mathcal{L}(\mathcal{U},\mathcal{Y})$ and $\mathbf{B} \in \mathcal{L}(\mathcal{Z},\mathcal{Y})$ be defined by

$$\langle \mathbf{A}u, w\rangle_{\mathcal{Y},\mathcal{Y}^*} = \int_\Gamma \rho(y)a(u(y),w(y);y)\mathrm{d}y \qquad \forall\, u \in \mathcal{U}, w \in \mathcal{Y}^* \tag{2.1.17a}$$

$$\langle \mathbf{B}z, w\rangle_{\mathcal{Y},\mathcal{Y}^*} = -\int_\Gamma \rho(y)b(w(y);z)\mathrm{d}y \qquad \forall\, z \in \mathcal{Z},\ w \in \mathcal{Y}^*. \tag{2.1.17b}$$

The definition of $\mathbf{A}$ and $\widehat{\mathbf{A}}$ implies

$$\langle \mathbf{A}u, w\rangle_{\mathcal{Y},\mathcal{Y}^*} = E[\langle \widehat{\mathbf{A}}u, w\rangle_{\mathcal{V}^*,\mathcal{V}}] \quad \forall\, u \in \mathcal{U}, w \in \mathcal{Y}^*.$$

2.1.3 Optimal Control of PDEs with Random Inputs

Consider the optimization problem (2.1.5). When the PDE coefficient, ϵ, in (2.1.3) is a random field, the solution, $u \in \mathcal{U}$, is also a random field. Hence, the map $y \mapsto j(u(y),z) : \Gamma \to \mathbb{R}$ for fixed $z \in \mathcal{Z}$ is a random variable. This randomness in the objective function is generally handled using "risk measures." Risk measures are operators which act on spaces of function with domain Γ and codomain $\mathbb{R}$ such as

$L^q_\rho(\Gamma)$ for $q \in [1,\infty)$ or $q = \infty$. Throughout this thesis, I will denote the risk measure as

$$\sigma : L^q_\rho(\Gamma) \to \mathbb{R}.$$

Assuming $y \in \Gamma \mapsto j(u(y), z)$ is a function in $L^q_\rho(\Gamma)$, the risk-averse optimization problem corresponding to the control of the linear elliptic PDE with uncertain coefficients, (2.1.9), is

$$\begin{array}{lll} \min\limits_{u\in\mathcal{U},\, z\in\mathcal{Z}} & J(u,z) \;:=\; \dfrac{1}{2}\sigma\Big(\|u-\bar{v}\|^2_{\mathcal{H}}\Big) \;+\; \dfrac{\alpha}{2}\|z\|^2_{\mathcal{Z}} & \\ \text{subject to} & \mathbf{A}u + \mathbf{B}z \;=\; 0. & \end{array} \tag{2.1.18}$$

Recall that for this example, $\mathcal{H} = \mathcal{Z} = L^2(D)$, and that $\mathbf{A}$, $\mathbf{B}$ are the operators defined in (2.1.17). Since the solution to the weak form (2.1.9) satisfies $u \in \mathcal{U} = L^p_\rho(\Gamma;\mathcal{V})$, the map

$$y \in \Gamma \mapsto \|u(y)-\bar{v}\|^2_{\mathcal{H}} \in L^{p/2}_\rho(\Gamma) \quad \text{or, equivalently,} \quad u \mapsto \|u-\bar{v}\|^2_{\mathcal{H}} : \mathcal{U} \to L^{p/2}_\rho(\Gamma).$$

Therefore, any risk measure used in (2.1.18) must have domain $L^{p/2}_\rho(\Gamma)$. Equivalently, given a risk measure

$$\sigma : L^q_\rho(\Gamma) \to \mathbb{R}$$

one requires the state space

$$\mathcal{U} = L^{2q}_\rho(\Gamma;\mathcal{V}) = L^{2q}_\rho(\Gamma; H^1_0(D)).$$

The corresponding image space for the operator (2.1.17) is

$$\mathcal{Y} = L^{2q}_\rho(\Gamma;\mathcal{V}^*) = L^{2q}_\rho(\Gamma; H^{-1}(D)).$$

Note that $\mathcal{Y}^* = L^{2q/(2q-1)}_\rho(\Gamma;\mathcal{V}) = L^{2q/(2q-1)}_\rho(\Gamma; H^1_0(D))$.

2.1.3.1 The Reduced Space Formulation and Differentiability

Optimization problem (2.1.18) can equivalently be written in the unconstrained reduced space form as

$$\min_{z\in\mathcal{Z}} \widehat{J}(z) := \sigma(\widehat{j}(z;y)) = \frac{1}{2}\sigma\Big(\|u(z)-\bar{v}\|^2_{\mathcal{H}}\Big) + \frac{\alpha}{2}\|z\|^2_{\mathcal{Z}}$$

where $u = u(z) \in \mathcal{U}$ solves $\mathbf{A}u + \mathbf{B}z = 0$. One can employ the adjoint calculus to compute derivatives of $\widehat{J}(z)$. Notice that the derivative of $\widehat{J}(z)$ requires multiple applications of the chain rule: the derivative of the risk measure, the derivative of the $\mathcal{H}$-norm, and the derivative of the solution to the state equation. From the deterministic optimization problem, the objective function,

$$j(v,z) = \frac{1}{2}\|v - \bar{v}\|_{\mathcal{H}}^2 + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2,$$

is continuously Fréchet differentiable with respect to both $v \in \mathcal{V}$ and $z \in \mathcal{Z}$. For general functions $f : \mathcal{V} \to \mathbb{R}$ which are Fréchet differentiable, it is unclear whether or not the mapping $u \in L_\rho^p(\Gamma;\mathcal{V}) \mapsto f(u)$ is also Fréchet differentiable. The quadratic objective function, j, is an example where Fréchet derivatives in $\mathcal{U} = L_\rho^p(\Gamma;\mathcal{V})$, $p = 2q$, are explicitly computable. In fact, there exists a positive constant $C > 0$ such that

$$\begin{aligned}\int_\Gamma \rho(y)\Big|\frac{1}{2}\|u(y) - \bar{v} + h(y)\|_{\mathcal{H}}^2 - \frac{1}{2}\|u(y) - \bar{v}\|_{\mathcal{H}}^2 - \langle u(y) - \bar{v}, h(y)\rangle_{\mathcal{H}}\Big|\mathrm{d}y \\ = \int_\Gamma \rho(y)\|h(y)\|_{\mathcal{H}}^2 \mathrm{d}y \leq \|h\|_{L_\rho^{2q}(\Gamma;\mathcal{H})}^2 \leq C\|h\|_{L_\rho^{2q}(\Gamma;\mathcal{V})}^2.\end{aligned}$$

This proves that for fixed $z \in \mathcal{Z}$ the mapping $u \in \mathcal{U} \mapsto j(u,z)$ is Fréchet differentiable and the derivative $j_u(u,z) \in \mathcal{L}(\mathcal{U}, L_\rho^q(\Gamma))$ for fixed $z \in \mathcal{Z}$. This result is extended to more general functions in the following proposition.

Proposition 2.1.1 *Let $f : \mathcal{V} \to \mathbb{R}$ be Fréchet differentiable and define $\beta : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ by the relationship*

$$|f(v+h) - f(v) - f'(v)h| = \beta(v,h)\|h\|_{\mathcal{V}}. \tag{2.1.19}$$

Consider $u \in L_\rho^p(\Gamma;\mathcal{V})$ with $p \in (1,\infty)$ or $p = \infty$ and assume the map

$$y \in \Gamma \mapsto f(u(y)) \in L_\rho^q(\Gamma)$$

for some $q \in [1,p)$. Then $f : L_\rho^p(\Gamma;\mathcal{V}) \to L_\rho^q(\Gamma)$ is Fréchet differentiable at u if

$$\lim_{\|h\|_{L_\rho^p(\Gamma;\mathcal{V})}\to 0} \|\beta(u,h)\|_{L_\rho^s(\Gamma)} = 0, \quad s := \frac{pq}{p-q} > 0. \tag{2.1.20}$$

Proof: Plugging u, $h \in L^p_\rho(\Gamma;\mathcal{V})$ into (2.1.19), taking the power q, and integrating over Γ yields

$$\int_\Gamma \rho(y)|f(u(y)+h(y))-f(u(y))-f'(u(y))h(y)|^q \mathrm{d}y = \int_\Gamma \rho(y)\beta(u(y),h(y))^q \|h(y)\|^q_{\mathcal{H}} \mathrm{d}y.$$

Applying Hölder's inequality to the right hand side gives

$$\begin{aligned}\int_\Gamma \rho(y)\beta(u(y),h(y))^q \|h(y)\|^q_{\mathcal{H}} \mathrm{d}y &\leq \left(\int_\Gamma \rho(y)\beta(u,h)^s \mathrm{d}y\right)^{q/s} \left(\int_\Gamma \rho(y)\|h\|^p_{\mathcal{H}} \mathrm{d}y\right)^{q/p} \\ &= \|\beta(u,h)\|^q_{L^s_\rho(\Gamma)} \|h\|^q_{L^p_\rho(\Gamma;\mathcal{H})}\end{aligned}$$

where s satisfies $\frac{1}{s/q} + \frac{1}{p/q} = 1$ (i.e. $s = \frac{pq}{p-q}$). Thus, if (2.1.20) holds, then

$$\lim_{\|h\|_{L^p_\rho(\Gamma;\mathcal{V})} \to 0} \|f(u+h) - f(u) - f'(u)h\|_{L^q_\rho(\Gamma)} = 0$$

and $f : L^p_\rho(\Gamma;\mathcal{V}) \to L^q_\rho(\Gamma)$ is Fréchet differentiable. □

Remark 2.1.2 *For fixed $z \in \mathcal{Z}$, the quadratic objective function,*

$$f(u) = j(u,z) = \frac{1}{2}\|v - \bar{v}\|^2_{\mathcal{H}} + \frac{\alpha}{2}\|z\|^2_{\mathcal{Z}},$$

satisfies the assumptions of Proposition 2.1.1 for $p = 2q$. Furthermore, $s = 2q$ and $\|\beta(v,h)\|_{L^{2q}_\rho(\Gamma)} = \|h\|_{L^{2q}_\rho(\Gamma;\mathcal{V})}$.

To conclude the formulation of optimization problem (2.1.18), I will assume that

$$\sigma : L^q_\rho(\Gamma) \to \mathbb{R} \quad \text{for } q \in [1,\infty).$$

By the uniform ellipticity assumption, (2.1.11), $u \in L^\infty_\rho(\Gamma;\mathcal{V})$ and one can choose the state space

$$\mathcal{U} := L^{2q}_\rho(\Gamma;\mathcal{V}).$$

Assuming the risk measure, σ, is Hadamard differentiable, the Fréchet differentiability of the map, $u \mapsto \|u - \bar{v}\|^2_{\mathcal{H}}$, guarantees that the map,

$$u \mapsto \sigma(\|u - \bar{v}\|^2_{\mathcal{H}}),$$

is Hadamard differentiable as a map from $\mathcal{U} := L_\rho^{2q}(\Gamma;\mathcal{V})$ into $\mathbb{R}$. Therefore, $J(u,z) = \frac{1}{2}\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$ is Hadamard differentiable with respect to $u \in \mathcal{U} = L_\rho^{2q}(\Gamma;\mathcal{V})$ and Fréchet differentiable with respect to $z \in \mathcal{Z}$.

2.1.3.2 The Adjoint Calculus

Recall $\sigma : L_\rho^q(\Gamma) \to \mathbb{R}$ for $q \in [1,\infty)$, $\mathcal{V} = H_0^1(D)$, $\mathcal{U} = L_\rho^{2q}(\Gamma;\mathcal{V})$, and $\mathcal{Y} = L_\rho^{2q}(\Gamma;\mathcal{V}^*)$. I will derive the adjoint calculus for this optimization problem by differentiating the Lagrangian functional, $L : \mathcal{U}\times\mathcal{Z}\times\mathcal{Y}^* \to \mathbb{R}$ given by

$$L(u,z,\lambda) := \frac{1}{2}\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 + \langle \mathbf{A}u + \mathbf{B}z, \lambda\rangle_{\mathcal{Y},\mathcal{Y}^*}.$$

The Lagrangian is Fréchet differentiable with respect to both $\lambda \in \mathcal{Y}^*$ and $z \in \mathcal{Z}$. On the other hand, L is Hadamard differentiable with respect to $u \in \mathcal{U}$ since the risk measure, σ, is Hadamard differentiable. Setting the Fréchet derivative of L with respect to the Lagrange multiplier, $\lambda \in \mathcal{Y}^*$, to zero returns the state equation (2.1.13). Differentiating L with respect to the state variable $u \in \mathcal{U}$ and setting the Hadamard derivative to zero results in the adjoint equation

$$\begin{aligned} 0 = \frac{\partial}{\partial u}L(u,z,\lambda)\delta u &= E[\nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)\langle u-\bar{v},\delta u\rangle_{\mathcal{H}}] + \langle \mathbf{A}\delta u,\lambda\rangle_{\mathcal{Y},\mathcal{Y}^*} \\ &= E[\nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)\langle u-\bar{v},\delta u\rangle_{\mathcal{H}}] + \langle \mathbf{A}^*\lambda,\delta u\rangle_{\mathcal{U}^*,\mathcal{U}} \end{aligned} \tag{2.1.21}$$

where $\mathbf{A}^* \in \mathcal{L}(\mathcal{Y}^*,\mathcal{U}^*)$. For the linear elliptic PDE (2.1.8), the following relationship holds

$$\langle u, \mathbf{A}^*w\rangle_{\mathcal{U},\mathcal{U}^*} = \langle \mathbf{A}u, w\rangle_{\mathcal{Y},\mathcal{Y}^*} = E[\langle \widehat{\mathbf{A}}u, w\rangle_{\mathcal{V}^*,\mathcal{V}}] = E[\langle u, \widehat{\mathbf{A}}^*w\rangle_{\mathcal{V},\mathcal{V}^*}]$$

for all $u \in \mathcal{U}$, $w \in \mathcal{Y}^*$. Therefore, the adjoint operators $\widehat{\mathbf{A}}(y)^* \in \mathcal{L}(\mathcal{V},\mathcal{V}^*)$ for almost all $y \in \Gamma$ and $\mathbf{A}^* \in \mathcal{L}(\mathcal{U},\mathcal{U}^*)$ satisfy

$$\langle u, \mathbf{A}^*w\rangle_{\mathcal{U},\mathcal{U}^*} = E[\langle u, \widehat{\mathbf{A}}^*w\rangle_{\mathcal{V},\mathcal{V}^*}] \quad \forall\, u \in \mathcal{U}, w \in \mathcal{Y}^*.$$

Substituting this equivalent parametrized expression for $\mathbf{A}$ in (2.1.21) yields

$$0 = E[\nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)\langle u-\bar{v},\delta u\rangle_{\mathcal{H}} + \langle \widehat{\mathbf{A}}^*\lambda,\delta u\rangle_{\mathcal{V}^*,\mathcal{V}}].$$

Now, since $\mathcal{H} = L^2(D) \subset \mathcal{V}^* = H^{-1}(D)$, the adjoint equation can be reformulated as

$$\begin{aligned} 0 &= E[\nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)\langle \mathbf{I}(u-\bar{v}), \delta u\rangle_{\mathcal{V}^*,\mathcal{V}} + \langle \widehat{\mathbf{A}}^*\lambda, \delta u\rangle_{\mathcal{V}^*,\mathcal{V}}] \\ &= E[\langle \nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)\mathbf{I}(u-\bar{v}) + \widehat{\mathbf{A}}^*\lambda, \delta u\rangle_{\mathcal{V}^*,\mathcal{V}}] \end{aligned}$$

where $\mathbf{I} \in \mathcal{L}(\mathcal{H}, \mathcal{V}^*)$ denotes the injection operator from $\mathcal{H}$ into $\mathcal{V}^*$.

As with the parametrized weak form, (2.1.15), it may be beneficial for the numerical solution of (2.3.1) to consider the parametrized adjoint equation

$$\widehat{\mathbf{A}}(y)^*\lambda(y) + \nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)\mathbf{I}(u(y)-\bar{v}) = 0 \tag{2.1.22}$$

where $u \in \mathcal{U} = L_\rho^{2q}(\Gamma;\mathcal{V})$. Since $\sigma \in L_\rho^q(\Gamma)$, it holds that $\nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2) \in L_\rho^{q/(q-1)}(\Gamma)$. Together with $\mathbf{I}(u-\bar{v}) \in L_\rho^{2q}(\Gamma;\mathcal{V}^*)$, one obtains

$$\nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)\mathbf{I}(u-\bar{v}) \in L_\rho^{2q/(2q-1)}(\Gamma;\mathcal{V}^*).$$

Moreover, the Lax-Milgram theorem ensures the existence of a unique solution, $\lambda(y) \in \mathcal{V}$ for almost every $y \in \Gamma$, to (2.1.22). Additionally, the uniform ellipticity assumption implies that $\|\widehat{\mathbf{A}}(\cdot)^{-*}\| \in L_\rho^\infty(\Gamma)$. Hence, the solution of the parametrized form of the adjoint equation (2.1.22) satisfies $\lambda \in \mathcal{Y}^* = L_\rho^{2q/(2q-1)}(\Gamma;\mathcal{V})$. For this linear elliptic test problem, the solution to the parametrized adjoint equation, (2.1.22), and the adjoint equation, (2.1.21), coincide due to uniqueness of solutions. In its strong form, the parametrized adjoint equation, (2.1.22), corresponds to the PDE

$$\begin{aligned} -\nabla\cdot(\epsilon(y,x)\nabla\lambda(y,x)) + \nabla\sigma(\|u-\bar{v}\|_{\mathcal{H}}^2)(u(y,x)-\bar{v}(x)) = 0 \quad & x \in D, \text{ a.e. in } \Gamma \\ \lambda(y,x) = 0 \quad & x \in \partial D, \text{ a.e. in } \Gamma. \end{aligned}$$

Finally, the gradient of $\widehat{J}(z)$ can be computed as

$$\nabla\widehat{J}(z) = \frac{\partial}{\partial z}L(u(z), z, \lambda(z))$$

where $u(z) = u \in \mathcal{U}$ solves the state equation (2.1.15) and $\lambda(z) = \lambda \in \mathcal{Y}^*$ solves the

adjoint equation (2.1.22). Thus, the gradient is given as follows

$$\begin{aligned}\nabla \widehat{J}(z)\delta z &= \alpha\langle z, \delta z\rangle_{\mathcal{Z}} + \langle \mathbf{B}\delta z, \lambda\rangle_{\mathcal{Y},\mathcal{Y}^*} \\ &= \langle \alpha z, \delta z\rangle_{\mathcal{Z}} + E[\langle \widehat{\mathbf{B}}\delta z, \lambda\rangle_{\mathcal{V}^*,\mathcal{V}}] \\ &= \langle \alpha z, \delta z\rangle_{\mathcal{Z}} + E[\langle \widehat{\mathbf{B}}^*\lambda, \delta z\rangle_{\mathcal{Z}}] \\ &= \langle \alpha z + E[\widehat{\mathbf{B}}^*\lambda], \delta z\rangle_{\mathcal{Z}}\end{aligned}$$

where for a.a. $y \in \Gamma$, $\widehat{\mathbf{B}}(y)^* \in \mathcal{L}(\mathcal{V}, \mathcal{Z})$ and

$$y \mapsto \widehat{\mathbf{B}}(y)^*\lambda(y) : \Gamma \to \mathcal{L}(\mathcal{Z}, L_\rho^{2q/(2q-1)}(\Gamma)),$$

since in our example $\|\widehat{\mathbf{B}}(\cdot)^*\| \in L_\rho^\infty(\Gamma)$.

The final equality above is due to Fubini's theorem [49]. This gives the following expression for the gradient

$$\nabla \widehat{J}(z) = \alpha z + E[\widehat{\mathbf{B}}^*\lambda].$$

2.1.3.3 Risk Measures

For demonstration purposes, I will now compute the derivative of $\widehat{J}(z)$ for a class of risk measures paying special attention to the integrability requirements of the risk measure. Members of the class under consideration have the form

$$\sigma(Y) = E[Y] + cE[\wp(Y - E[Y])]$$

where $\wp : \mathbb{R} \to \mathbb{R}$ is assumed to be differentiable. Particular members of this class of risk measures are the "mean plus moment" risk measures and, with some care, the derivations presented here apply to the "mean plus semi-deviation" risk measure. Computing the derivative of σ yields

$$\begin{aligned}E[\nabla\sigma(Y)\eta] &= E[\eta] + cE[\wp'(Y - E[Y])(\eta - E[\eta])] \\ &= E[\eta] + cE[\wp'(Y - E[Y])\eta] - cE[\wp'(Y - E[Y])]E[\eta] \\ &= E\Big[\eta + c\Big(\wp'(Y - E[Y]) - E\big[\wp'(Y - E[Y])\big]\Big)\eta\Big]\end{aligned}$$

which gives the gradient of σ as

$$\nabla\sigma(Y) = 1 + c\Big(\wp'(Y - E[Y]) - E\big[\wp'(Y - E[Y])\big]\Big).$$

As a first example, I will focus on the mean plus variance risk measure. In this case, $\wp(Y) = \frac{1}{2}Y^2$ and σ has domain $L^2_\rho(\Gamma)$. Consequently, the state space is $\mathcal{U} = L^4_\rho(\Gamma;\mathcal{V})$ so that for $u \in \mathcal{U}$ the map $y \mapsto \|u(y) - \bar{v}\|^2_{\mathcal{H}} \in L^2_\rho(\Gamma)$. For this risk measure, $\wp'(Y) = Y$ and the derivative of σ has the particularly simple form

$$\nabla\sigma(Y) = 1 + c(Y - E[Y]) \in L^2_\rho(\Gamma).$$

In the case of mean plus semi-deviation, $\wp(Y) = [Y]_+ = \max\{Y, 0\}$, and $\sigma : L^1_\rho(\Gamma) \to \mathbb{R}$. In this case the state space is $\mathcal{U} = L^2_\rho(\Gamma;\mathcal{V})$. Clearly, $\wp$ is not differentiable at $Y = 0$, but is continuously differentiable everywhere in $\mathbb{R} \setminus \{0\}$. In fact, $\wp'(Y) \equiv 1$ if $Y > 0$ and $\wp'(Y) \equiv 0$ if $Y < 0$. Therefore, if $Y \in L^1_\rho(\Gamma)$ and $Y \neq E[Y]$ almost everywhere in Γ, then

$$\nabla\sigma(Y) = \begin{cases} 1 + c(1 - \Pr(Y > E[Y])) & \text{if } Y > E[Y] \\ 1 - c\Pr(Y > E[Y]) & \text{if } Y < E[Y]. \end{cases}$$

Clearly $\nabla\sigma(Y) \in L^\infty_\rho(\Gamma)$ for all $Y \in L^1_\rho(\Gamma)$ such that $Y \neq E[Y]$ almost everywhere in Γ since $|\nabla\sigma(Y)| \leq 1 + c$.

2.2 General Problem Formulation

Let $\mathcal{V}$ and $\mathcal{Z}$ be reflexive Banach spaces, and let $\mathcal{W}$ be a Banach space. Throughout this thesis, $\mathcal{V}$ will denote the deterministic state space and $\mathcal{Z}$ will denote the control space. As with the test problem, the stochastic program considered in this thesis is an extension of the deterministic equality constrained problem

$$\min_{v\in\mathcal{V},\, z\in\mathcal{Z}} j(v, z) \quad \text{subject to} \quad \tilde{e}(v, z) = 0, \tag{2.2.1}$$

where the objective function, $j : \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$, corresponds to the "cost" associated with a given state $v \in \mathcal{V}$ and a given control $z \in \mathcal{Z}$ and the equality constraint,

$\tilde{e} : \mathcal{V} \times \mathcal{Z} \to \mathcal{W}$, represents the governing dynamics (PDE or system of PDEs). The stochastic variant of (2.2.1) is generated by adding uncertainty or randomness to the state operator, $\tilde{e}$. Throughout this thesis, the control variable will be deterministic and thus adding uncertainty to $\tilde{e}$ induces randomness in the state variable. The stochastic state space will be denoted $\mathcal{U}$. As seen in the test problem, the uncertainty in the state variable induces uncertainty in the objective function which is typically handled with risk measures. The risk-averse optimization problem corresponding to (2.2.1) is

$$\min_{u \in \mathcal{U},\, z \in \mathcal{Z}} \sigma\big(j(u,z)\big) \quad \text{subject to} \quad e(u,z) = 0, \tag{2.2.2}$$

where σ is a risk measure, and $e : \mathcal{U} \times \mathcal{Z} \to \mathcal{Y}$ for some real Banach space $\mathcal{Y}$ denotes the stochastic state equation. This goal of this section is to make clear the functional analytic framework for (2.2.2) by making assumptions on function spaces as well as the objective function and state equation. The assumption presented here ensure well posedness of (2.2.2).

2.2.1 The State Equation

To begin, I will present assumptions on that state equation to ensure (2.2.2) is well posed. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space where Ω is the set of outcomes, $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra of events, and $P : \mathcal{F} \to [0,1]$ is a probability measure. When uncertainty is added to the deterministic state operator $\tilde{e}$, I will denote the stochastic variant as $\tilde{e}(\omega) : \mathcal{V} \times \mathcal{Z} \to \mathcal{W}$ for almost every $\omega \in \Omega$. My first assumption is a common assumption in the literature and is known as the "Finite Dimensional Noise Assumption." Finite dimensional noise assumes the operator, $\tilde{e}(\omega)$, has a finite dependence on the random variable $\omega \in \Omega$. This finite dependence will facilitate the numerical solution of the optimization problem (2.2.2).

Assumption 2.2.1 (Finite Dimensional Noise) *There exists a vector of random*

variables, $Y = [Y_1, \ldots, Y_M] : \Omega \to \Gamma \subseteq \mathbb{R}^M$*, such that*

$$\tilde{e}(\omega) \equiv \tilde{e}(Y(\omega))$$

where $Y_i : \Omega \to \Gamma_i \subseteq \mathbb{R}$ *are independent random variables with image space* $\Gamma_i = [a_i, b_i]$*,* $a_i < b_i$*, and Lebesgue density* $\rho_i : \Gamma_i \to \mathbb{R}$ *for* $i = 1, \ldots, M$*. The joint image space of* Y *is* $\Gamma = \prod_{i=1}^{M} \Gamma_i$ *and the joint density is* $\rho = \prod_{i=1}^{M} \rho_i$*.*

This assumption allows for the change of variables

$$\tilde{e}(y)(v, z) = \tilde{e}(v, z; y) = 0 \quad \forall\, y \in \Gamma. \tag{2.2.3}$$

Furthermore, this permits the use of finite sampling and polynomial approximation schemes in solving (2.2.3) (c.f. [9, 10, 11]). The Karhunen-Loéve (KL) expansion of a random field is an infinite dimension extension of the singular value decomposition (SVD) [64, 71, 103]. Truncating the KL expansion gives a finite noise approximation of the random field and this truncation is often used to satisfy Assumption 2.2.1. I will discuss this expansion technique in more detail in Section 2.4.

The parametrized state equation (2.2.3) will be utilized when solving (2.2.2) numerically, but will not be used in the analysis of (2.2.2). Let $\mathcal{Y}$ be a space of functions on Γ with values in $\mathcal{W}$. Furthermore, let $\mathcal{U}$ be a space of functions on Γ with values in $\mathcal{V}$. The state equation used for analysis is

$$e(u, z) = 0 \tag{2.2.4}$$

where $e : \mathcal{U} \times \mathcal{Z} \to \mathcal{Y}$. To make the spaces $\mathcal{Y}$ and $\mathcal{U}$ concrete and to make the definition of risk measures and other quantities coherent, I will assume integrability of the functions in $\mathcal{U}$ and $\mathcal{Y}$ with respect to $y \in \Gamma$. This integrability will ensure that the solution to the state equation and elements in $\mathcal{Y}$ have certain statistical moments.

Assumption 2.2.2 (Integrability of the State) *The state space for* (2.2.4) *is the Bochner space*

$$\mathcal{U} := L^p_\rho(\Gamma; \mathcal{V}) \quad \textit{for } p \in (1, \infty) \textit{ or } p = \infty$$

and the parametrized state operator satisfies

$$(u, z, y) \mapsto \tilde{e}(u(y), z; y) : \mathcal{U} \times \mathcal{Z} \times \Gamma \to \mathcal{Y} := L^s_\rho(\Gamma; \mathcal{W}) \tag{2.2.5}$$

for some $s \in [1, \infty)$.

Remark 2.2.3 *Equation 2.2.5 in Assumption 2.2.2 implies*

$$y \mapsto \langle \lambda(y), \tilde{e}(u(y), z; y) \rangle_{\mathcal{W}^*, \mathcal{W}} : \Gamma \to L^1_\rho(\Gamma)$$

for all $\lambda \in \mathcal{Y}^* := L^{s^*}_\rho(\Gamma; \mathcal{W})$ *with* $\frac{1}{s} + \frac{1}{s^*} = 1$, $u \in \mathcal{U}$, *and* $z \in \mathcal{Z}$ *or, equivalently,*

$$(u, z, \lambda) \mapsto \langle \lambda, \tilde{e}(u, z; \cdot) \rangle_{\mathcal{W}^*, \mathcal{W}} : \mathcal{U} \times \mathcal{Z} \times \mathcal{Y}^* \to L^1_\rho(\Gamma). \tag{2.2.6}$$

The state equation, (2.2.4), is thus defined by the relationship

$$\langle \lambda, e(u, z) \rangle_{\mathcal{Y}^*, \mathcal{Y}} = \int_\Gamma \rho(y) \langle \lambda(y), \tilde{e}(u(y), z; y) \rangle_{\mathcal{W}^*, \mathcal{W}} \mathrm{d}y = E[\langle \lambda, \tilde{e}(u, z; \cdot) \rangle_{\mathcal{W}^*, \mathcal{W}}] \tag{2.2.7}$$

for all $\lambda \in \mathcal{Y}^*$.

To ensure that optimization problem (2.2.2) is well defined, for each $z \in \mathcal{Z}$, there must exist $u(z) = u \in \mathcal{U}$ such that (2.2.4) is satisfied. To simplify presentation, I will use the following notation to denote partial Fréchet derivatives

$$e_u(u, z) := \frac{\partial}{\partial u} e(u, z) \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$$

and similarly for derivatives with respect to $z \in \mathcal{Z}$. I will employ similar notation for derivatives of the objective function j.

Assumption 2.2.4 (Existence of Solution Mapping)

- *For all* $z \in \mathcal{Z}$ *there exists a unique* $u \in \mathcal{U}$ *such that* $e(u, z) = 0$*;*
- *There exists an open set* $\Sigma \subset \mathcal{U} \times \mathcal{Z}$ *with*

$$\mathcal{S} := \{(u, z) \in \mathcal{U} \times \mathcal{Z} \ : \ e(u, z) = 0\} \subset \Sigma$$

such that $e(u, z)$ *are Fréchet differentiable on* Σ*;*

- *The inverse* $e_u(u,z)^{-1} \in \mathcal{L}(\mathcal{Y},\mathcal{U})$ *exists for all* $(u,z) \in \mathcal{S}$.

Assumptions 2.2.4 ensure that the Implicit Function Theorem (Theorem 1.41 in [63]) holds for each $z \in \mathcal{Z}$. In addition to Assumption 2.2.4, I will require differentiability of the parametrized state operator, $\tilde{e}$.

Assumption 2.2.5 *Consider the parametrized state operator,* $\tilde{e}(y) : \mathcal{V} \times \mathcal{Z} \to \mathcal{W}$ *for fixed* $y \in \Gamma$. *Then the mapping*

$$(u,z,y) \mapsto \tilde{e}(u,z;y) : \mathcal{U} \times \mathcal{Z} \times \Gamma \to \mathcal{Y}$$

is Fréchet differentiable with respect to $(u,z) \in \Sigma$. *Furthermore, the partial derivative,* $\tilde{e}_u$, *has a bounded inverse*

$$\tilde{e}_u(u(y),z;y)^{-1} \in \mathcal{L}(\mathcal{W},\mathcal{V}) \quad \forall\ (u,z) \in \Sigma,\ \textit{a.e. in}\ \Gamma$$

and the partial derivative, $\tilde{e}_z$, *satisfies*

$$\begin{aligned} E[\langle \lambda, \tilde{e}_z(u,z;\cdot)\delta z\rangle_{\mathcal{W}^*,\mathcal{W}}] &= E[\langle \tilde{e}_z^*(u,z;\cdot)\lambda, \delta z\rangle_{\mathcal{Z}^*,\mathcal{Z}}] \\ &= \langle E[\tilde{e}_z^*(u,z;\cdot)\lambda], \delta z\rangle_{\mathcal{Z}^*,\mathcal{Z}}. \end{aligned} \tag{2.2.8}$$

Assumption 2.2.5 ensures the Fréchet derivatives of the state operator, e, can be written in terms of the Fréchet derivatives of the parametrized state operator, $\tilde{e}$,

$$\langle \lambda, e_u(u,z)\delta u\rangle_{\mathcal{Y}^*,\mathcal{Y}} = \int_\Gamma \rho(y)\langle \lambda(y), \tilde{e}_u(u(y),z;y)\delta u\rangle_{\mathcal{W}^*,\mathcal{W}} \mathrm{d}y$$

and

$$\langle \lambda, e_z(u,z)\delta z\rangle_{\mathcal{Y}^*,\mathcal{Y}} = \int_\Gamma \rho(y)\langle \lambda(y), \tilde{e}_z(u(y),z;y)\delta z\rangle_{\mathcal{W}^*,\mathcal{W}} \mathrm{d}y.$$

Moreover, equation 2.2.8 is a condition similar to the conclusion of Fubini's Theorem [49].

2.2.2 The Objective Function

Assumption 2.2.4 ensures the existence of a unique solution, $u \in \mathcal{U}$, to the state equation (2.2.4) for all $z \in \mathcal{Z}$. Furthermore, this assumption guarantees the well-posedness of the reduced formulation

$$\min_{z \in \mathcal{Z}} \widehat{J}(z) := \sigma(\widehat{j}(z;y)) \tag{2.2.9}$$

for appropriate risk measures σ. Under the finite noise assumption, the reduced objective function is

$$y \mapsto \widehat{j}(z;y) := j(u(y;z),z) : \Gamma \to \mathbb{R}.$$

To make concrete the notion of risk measure, I will require that the map $y \mapsto \widehat{j}(z;y)$ is sufficiently integrable.

Assumption 2.2.6 (Integrability of the Objective Function) *For $u \in \mathcal{U} = L^p_\rho(\Gamma;\mathcal{V})$, the mapping $y \in \Gamma \mapsto j(u(y),z) \in L^q_\rho(\Gamma)$ for $q \in [1,p)$.*

Remark 2.2.7 *Assumption 2.2.6 is equivalent to*

$$(u,z) \mapsto j(u,z) : \mathcal{U} \times \mathcal{Z} \to L^q_\rho(\Gamma).$$

Of particular interest to this thesis are derivative based algorithms. These algorithms require differentiability of $j(v,z)$. In addition to differentiability of $j(v,z)$, I will require differentiability of the risk measure, $\sigma(Y)$. Since $\widehat{J}(z) = \sigma(\widehat{j}(z;y))$ is a composite function, the chain rule must hold for the derivatives of $\sigma(Y)$.

Assumption 2.2.8 (Differentiability of the Objective Function)

- *The deterministic objective function, $j : \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$, is Fréchet differentiable with respect to $v \in \mathcal{V}$ and for fixed $z \in \mathcal{Z}$ the functional $\beta_1 : \mathcal{V} \times \mathcal{Z} \times \mathcal{V} \to \mathbb{R}$ defined by*

$$|j(v+h,z) - j(v,z) - j_v(v,z)h| = \beta_1(v,z,h)\|h\|_{\mathcal{V}}$$

satisfies

$$\lim_{\|h\|_{L^p_\rho(\Gamma;\mathcal{V})}\to 0} \|\beta_1(u,z,h)\|_{L^q_\rho(\Gamma)} = 0 \quad \textit{where} \quad \theta := \frac{pq}{p-q} > 0$$

for each $u \in \mathcal{U} = L^p_\rho(\Gamma;\mathcal{V})$.

- *The deterministic objective function,* j*, is also Fréchet differentiable with respect to* $z \in \mathcal{Z}$ *and for fixed* $v \in \mathcal{V}$ *the functional* $\beta_2 : \mathcal{V} \times \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ *defined by*

$$|j(v,z+h) - j(v,z) - j_z(v,z)h| = \beta_2(v,z,h)\|h\|_{\mathcal{Z}}$$

satisfies, for fixed $u \in \mathcal{U}$*,*

$$\lim_{\|h\|_{\mathcal{Z}}\to 0} \|\beta_2(u,z,h)\|_{L^q_\rho(\Gamma)} = 0 \quad \forall\, z \in \mathcal{Z}.$$

- *The partial derivative,* $j_z(u,z)$ *for* $(u,z) \in \mathcal{U} \times \mathcal{Z}$*, satisfies*

$$E[j_z(u,z)\delta z] = E[j_z(u,z)]\delta z \quad \forall\, \delta z \in \mathcal{Z}.$$

Assumptions 2.2.8 and Proposition 2.1.1 ensure that the deterministic objective function is Fréchet differentiable with respect to $u \in \mathcal{U}$ and $z \in \mathcal{Z}$. Furthermore, Assumption 2.2.6 implies that

$$j_u(u,z) \in \mathcal{L}(\mathcal{U}, L^q_\rho(\Gamma)) \quad \text{and} \quad j_z(u,z) \in \mathcal{L}(\mathcal{Z}, L^q_\rho(\Gamma)).$$

The final condition in Assumption 2.2.8 is related to the conclusions of Fubini's Theorem [49].

Since for $u \in \mathcal{U}$ and $z \in \mathcal{Z}$ the function $y \mapsto j(u(y),z) \in L^q_\rho(\Gamma)$, our risk measure must satisfy

$$\sigma \in L^q_\rho(\Gamma).$$

Of course, since $\int_\Gamma \rho(y)dy = 1$, any risk measure $\sigma \in L^r_\rho(\Gamma)$ with $r \geq q$ satisfies $\sigma \in L^q_\rho(\Gamma)$.

To guarantee the existence of derivatives of the reduced objective function, $\widehat{J}(z)$, I will require the following assumption on the risk measure.

Assumption 2.2.9 (Differentiability of Risk Measure) *The risk measure,* $\sigma : L^q_\rho(\Gamma) \to \mathbb{R}$*, is Hadamard differentiable.*

Assumption 2.2.9 implies that $\widehat{J}(z)$ is at least Fréchet differentiable and a gradient exists if $\mathcal{Z}$ is a Hilbert space. Note that Hadamard differentiability is required since Hadamard differentiability is the weakest form of differentiability for which the chain rule for computing derivatives of composite functions holds. For a review of the various notions of differentiability in linear topological spaces see [6, 7, 107]. Furthermore, to reinforce notation, Table 2.1 contains a description of the numerous powers used throughout this section in the definition of the Lebesgue and Bochner spaces.

Power	Range	Description
p	$[1,\infty) \cup \{\infty\}$	Integrability of state space, $\mathcal{U}$
q		Integrability of $(u,z) \mapsto j(u,z)$ for $(u,z) \in \mathcal{U} \times \mathcal{Z}$ and domain of risk measure
s	$[1,\infty) \cup \{\infty\}$	Integrability of codomain of the state operator, $\mathcal{Y}$

Table 2.1: Description of the powers for the Lebesgue and Bochner spaces used in the assumptions on the objective function and state equation

To ensure the existence of at least one minimizer of (2.2.9), I require that $\widehat{J}(z)$ is weakly lower semi-continuous and satisfies the infinity property [63].

Assumption 2.2.10 (Infinity Property) *For all sequences,* $\{z_n\} \subset \mathcal{Z}$ *such that* $\|z_n\|_{\mathcal{Z}} \to +\infty$*, the objective function satisfies* $\widehat{J}(z_n) \to +\infty$*.*

In addition, a unique minimizer is guaranteed when $\widehat{J}(z)$ is uniformly convex.

An extremely useful class of deterministic objective functions, $j : \mathcal{V} \times \mathcal{Z} \to \mathbb{R}$, is the class of "separable" objective functions

$$j(v,z) = j_1(v) + j_2(z)$$

where $j_1 : \mathcal{V} \to \mathbb{R}$ is convex, $j_2 : \mathcal{Z} \to \mathbb{R}$ is uniformly convex, and $j(v, z)$ satisfies Assumptions 2.2.8 and 2.2.6. This is the case for quadratic control or least squares type problems with regularization. For these separable objective functions, the implicitly constrained optimization problem (2.2.9) can be written as

$$\min_{z \in \mathcal{Z}} \widehat{J}(z) := \sigma\Big(j_1(u(y; z)) + j_2(z)\Big).$$

The class of separable objective functions warrants a few assumptions on the risk measure, $\sigma(Y)$. Note $j_2(z)$ does not depend on $y \in \Gamma$ and thus should not contribute to the risk associated with the state $u \in \mathcal{U}$. Furthermore, since $j(v, z) = j_1(v) + j_2(z)$ is convex, $\sigma(Y)$ should maintain this convexity.

Assumption 2.2.11 (Coherent Risk Measure) *Assume $\sigma : L^q_\rho(\Gamma) \to \mathbb{R}$ satisfies*

- ***Convexity:*** *For all $Y_1,\ Y_2 \in L^q_\rho(\Gamma)$ and $\lambda \in [0, 1]$,*

$$\sigma(\lambda Y_1 + (1 - \lambda)Y_2) \leq \lambda\sigma(Y_1) + (1 - \lambda)\sigma(Y_2);$$

- ***Monotonicity:*** *For all $Y_1,\ Y_2 \in L^q_\rho(\Gamma)$ such that $Y_1 \leq Y_2$ a.e.,*

$$\sigma(Y_1) \leq \sigma(Y_2);$$

- ***Translation Equivariance:*** *For all $Y \in L^q_\rho(\Gamma)$ and $c \in \mathbb{R}$,*

$$\sigma(Y + c) = \sigma(Y) + c;$$

- ***Positive Homogeneity:*** *For all $c > 0$ and $Y \in L^q_\rho(\Gamma)$,*

$$\sigma(cY) = c\sigma(Y).$$

Risk measures which satisfy Assumptions 2.2.11 are called *coherent* in the sense of Artzner, Delbaen, Eber, and Heath [5]. These assumptions imply the following form for the separable objective function

$$\widehat{J}(z) = \sigma\Big(j_1(u(y; z))\Big) + j_2(z).$$

Furthermore, $\widehat{J}(z)$ is uniformly convex whenever $\sigma(j_1(u(y; z)))$ is convex.

Remark 2.2.12 *Two particularly important coherent risk measures are the mean plus semi-deviation risk measure and the conditional value-at-risk (CVaR) risk measure. The mean plus semi-deviation risk measure of order q is the risk measure $\sigma_r : L^q_\rho(\Gamma) \to \mathbb{R}$ defined by*

$$\sigma_r(Y) := E[Y] + cE[[Y - E[Y]]^q_+]^{1/q}$$

for $0 \leq c \leq 1$ where $[x]_+ = \max\{x, 0\}$, [100]. On the other hand, the CVaR risk measure is $\sigma : L^1_\rho(\Gamma) \to \mathbb{R}$ defined by

$$\sigma_{CVaR}(Y) := \min_{t \in \mathbb{R}} \left\{ t + cE[[Y - t]_+] \right\}$$

for $c > 1$ [117]. These two coherent risk measures fall into a more general class of risk measures defined by the auxiliary function $\widehat{\sigma}_r : \mathbb{R} \times L^q_\rho(\Gamma) \to \mathbb{R}$ defined by

$$\widehat{\sigma}_r(t, Y) := t + cE[[Y - t]^q_+]^{1/q}.$$

That is, the mean plus semi-deviation and CVaR risk measures can be expressed as

$$\sigma_r(Y) = \widehat{\sigma}_r(E[Y], Y) \quad \textit{and} \quad \sigma_{CVaR}(Y) = \min_{t \in \mathbb{R}} \widehat{\sigma}_1(t, Y),$$

respectively. Furthermore, these two coherent risk measures are Hadamard differentiable, i.e. they satisfy both Assumption 2.2.8 and Assumption 2.2.11.

2.3 The Adjoint Calculus

I will now derive an adjoint calculus for computing the derivative of $\widehat{J}(z)$ (c.f. see Chapter 1.6 of [63] for a detailed discussion of adjoints). To clarify notation, I will denote the derivative of $\widehat{J}(z)$ as $\widehat{J}'(z) \in \mathcal{Z}^*$. For the risk measure σ, the derivative $\sigma'(Y)$ can be associated with $\nabla\sigma(Y) \in L^{q^*}_\rho(\Gamma)$ where $1 = \frac{1}{q} + \frac{1}{q^*}$ and $\sigma'(Y)s = E[\nabla\sigma(Y)s]$ for all $s \in L^q_\rho(\Gamma)$.

To derive the adjoint calculus, consider the Lagrangian functional

$$L : \mathcal{U} \times \mathcal{Z} \times \mathcal{Y}^* \to \mathbb{R}$$

defined by

$$L(u,z,\lambda) := \sigma(j(u,z)) + \langle \lambda, e(u,z)\rangle_{\mathcal{Y}^*,\mathcal{Y}}.$$

Note that the Lagrangian is at least Hadamard differentiable with respect to $u \in \mathcal{U}$ and $z \in \mathcal{Z}$ by Assumptions 2.2.8, 2.2.9, and 2.2.4. Furthermore, L is Fréchet differentiable with respect to the Lagrange multiplier $\lambda \in \mathcal{Y}^*$ by linearity. The Lagrangian functional is often used as a local lower support function for proving optimality conditions and satisfies the condition

$$\langle \widehat{J}'(z), \delta z\rangle_{\mathcal{Z}^*,\mathcal{Z}} = \langle L_z(u,z,\lambda), \delta z\rangle_{\mathcal{Z}^*,\mathcal{Z}}$$

whenever $u \in \mathcal{U}$ and $\lambda \in \mathcal{Y}^*$ satisfy

$$L_u(u,z,\lambda) = 0 \quad \text{and} \quad L_\lambda(u,z,\lambda) = 0.$$

Differentiating L with respect to the Lagrange multiplier $\lambda \in \mathcal{Y}^*$ and setting the derivative equal to zero returns the state equation (2.2.4). On the other hand, setting the derivative of L with respect to $u \in \mathcal{U}$ to zero yields the adjoint equation

$$\begin{aligned} 0 = L_u(u,z,\lambda)\delta u &= E[\nabla\sigma(j(u,z))j_u(u,z)\delta u] + \langle \lambda, e_u(u,z)\delta u\rangle_{\mathcal{Y}^*,\mathcal{Y}} \\ &= E[\nabla\sigma(j(u,z))j_u(u,z)\delta u] + \langle e_u(u,z)^*\lambda, \delta u\rangle_{\mathcal{U}^*,\mathcal{U}} \end{aligned}$$

for all $\delta u \in \mathcal{U}$ where $j_u(u,z) \in \mathcal{L}(\mathcal{U}, L^q_\rho(\Gamma))$ and $e_u(u,z)^* \in \mathcal{L}(\mathcal{Y}^*,\mathcal{U}^*)$ denotes the adjoint of $e_u(u,z)$. Now, unraveling the duality pairing, $\langle \cdot, \cdot \rangle_{\mathcal{U}^*,\mathcal{U}}$, and employing Assumption 2.2.5, the adjoint equation can be written as

$$0 = E[\langle \nabla\sigma(j(u,z))j_u(u,z) + \tilde{e}_u(u,z;\cdot)^*\lambda, \delta u\rangle_{\mathcal{V}^*,\mathcal{V}}] \quad \forall\, \delta u \in \mathcal{U}. \tag{2.3.1}$$

As in the test case, it may be beneficial to consider the parametrized adjoint equation

$$\tilde{e}_u(u(y),z;y)^*\lambda(y) + \nabla\sigma(j(u,z))j_u(u(y),z) = 0 \quad \text{a.e. } \in \Gamma. \tag{2.3.2}$$

Assumption 2.2.5 ensures that $\tilde{e}_u(u,z;y)$ has a bounded inverse almost every where in Γ; therefore, (2.3.2) has a unique solution, $\lambda(y) \in \mathcal{W}^*$ for almost all $y \in \Gamma$. By Assumption 2.2.4, the solutions to (2.3.2) and (2.3.1) coincide.

Finally, differentiating the Lagrangian with respect to $z \in \mathcal{Z}$ and applying Assumptions 2.2.8 and 2.2.5 yields

$$\begin{aligned} L_z(u,z,\lambda)\delta z &= E[\nabla\sigma(j(u,z))j_z(u,z)\delta z] + \langle \lambda, e_z(u,z)\delta z\rangle_{\mathcal{Y}^*,\mathcal{Y}} \\ &= \langle E[\nabla\sigma(j(u,z))j_z(u,z) + \tilde{e}_z^*(u,z;\cdot)\lambda], \delta z\rangle_{\mathcal{Z}^*,\mathcal{Z}} \end{aligned}$$

for all $\delta z \in \mathcal{Z}$ where $j_z(u,z) \in \mathcal{L}(\mathcal{Z}, L_\rho^q(\Gamma))$ and $e_z^*(u,z) \in \mathcal{L}(\mathcal{Y}^*, \mathcal{Z}^*)$ denotes the adjoint of $e_z(u,z)$. Therefore, for fixed $z \in \mathcal{Z}$, if $u(z) = u \in \mathcal{U}$ solves (2.2.4) and $\lambda(z) = \lambda \in \mathcal{Y}^*$ solves (2.3.2), then the derivative of $\widehat{J}$ is

$$\widehat{J}'(z) = E[\nabla\sigma(j(u,z))j_z(u,z) + \tilde{e}_z^*(u,z;\cdot)\lambda]. \tag{2.3.3}$$

Remark 2.3.1 *If σ is a coherent risk measure (i.e. satisfies Assumptions 2.2.11) and j is a separable objective function with*

$$j(v,z) = j_1(v) + j_2(z),$$

then (2.3.3) *can be simplified to*

$$\widehat{J}'(z) = E\Big[\tilde{e}_z^*(u,z;\cdot)\lambda\Big] + j_2'(z) \tag{2.3.4}$$

where $u = u(z) \in \mathcal{U}$ solves (2.2.4) *and $\lambda = \lambda(z) \in \mathcal{W}^*$ solves*

$$\tilde{e}_u^*(u(y),z;y)\lambda + \nabla\sigma(j_1(u))j_1'(u(y)) = 0. \tag{2.3.5}$$

Now, returning to the test problem (2.1.1), the objective function j is separable so Remark 2.3.1 applies. Recall the objective function is given as $j(v,z) = j_1(v) + j_2(z)$ where

$$j_1(v) = \frac{1}{2}\|\mathbf{Q}v - \bar{q}\|_{\mathcal{H}}^2 \quad \text{and} \quad j_2(z) = \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2.$$

The adjoint equation (2.3.5) for this specific objective function is

$$\widehat{\mathbf{A}}(y)^*\lambda(y) + \nabla\sigma(\|\mathbf{Q}u - \bar{q}\|_{\mathcal{H}}^2)\mathbf{Q}^*(\mathbf{Q}u(y) - \bar{q}) = 0$$

where $\mathbf{Q}^* \in \mathcal{L}(\mathcal{H}^*, \mathcal{V}^*)$ denotes the adjoint operator associated with $\mathbf{Q}$ and $u = u(z) \in \mathcal{U}$ solves the state equation for a given $z \in \mathcal{Z}$. On the other hand, the derivative of

$\widehat{J}(z)$ is handled in a similar fashion to that of (2.3.4). First notice that the derivative of the deterministic objective function, $j(v,z)$, with respect to $z \in \mathcal{Z}$ in the direction $s \in \mathcal{Z}$ satisfies

$$\langle j_z(v,z), s\rangle_{\mathcal{Z}^*,\mathcal{Z}} = \langle \alpha z, s\rangle_{\mathcal{Z}} = \langle \alpha \mathbf{R} z, s\rangle_{\mathcal{Z}^*,\mathcal{Z}}$$

where $\mathbf{R} \in \mathcal{L}(\mathcal{Z}, \mathcal{Z}^*)$ is the unique operator satisfying $\langle \mathbf{R}z, s\rangle_{\mathcal{Z}^*,\mathcal{Z}} = \langle z, s\rangle_{\mathcal{Z}}$. This gives rise to the following expression for the derivative (2.3.4)

$$\widehat{J}'(z) = \alpha \mathbf{R} z + E[\widehat{\mathbf{B}}^* \lambda]$$

where $\lambda \in \mathcal{Y}^*$ solves the adjoint equation. Note that since $\mathcal{Z}$ is assumed to be a Hilbert space and $\widehat{J}(z)$ is at least Hadamard differentiable by Assumption 2.2.8, Riesz Representation Theorem (e.g. see Theorem 1.4 in [63]) ensures the existence of a representer for $\widehat{J}'(z)$ in $\mathcal{Z}$ (i.e. the gradient). The above calculations give the following expression for the gradient of $\widehat{J}(z)$

$$\nabla \widehat{J}(z) = \alpha z + w$$

where $w = w(z) \in \mathcal{Z}$ solves

$$\mathbf{R} w = E\Big[\nabla \sigma(\|\mathbf{Q}u - \bar{q}\|^2_{\mathcal{H}}) \mathbf{B}^* \lambda\Big],$$

$u = u(z) \in \mathcal{U}$ solves the state equation, and $\lambda = \lambda(z) \in \mathcal{Y}^*$ solves the adjoint equation.

2.4 The Karhunen-Loéve Expansion

In this section, I will discuss a common technique for satisfying the finite noise assumption, Assumption 2.2.1. This technique is known as the Karhunen-Loéve (KL) expansion [64, 71]. The KL expansion gives an infinite series representation of a given random field. Often, this series representation is truncated and the original random field is replaced by a partial sum.

Now, let $\epsilon : \Omega \times D \to \mathbb{R}$ be a random field on the probability space $(\Omega, \mathcal{F}, P)$ with finite second order moments, $\epsilon \in L^2_P(\Omega; L^2(D))$. Associated with ϵ is the covariance function

$$C_\epsilon(x, \chi) := E\Big[(\epsilon(\cdot, x) - E[\epsilon(\cdot, x)])(\epsilon(\cdot, \chi) - E[\epsilon(\cdot, \chi)])\Big]$$

where $E : L^1_P(\Gamma) \to \mathbb{R}$ denotes the expected value operator $E[X] = \int_\Omega X(\omega)\mathrm{d}P(\omega)$. The covariance function, $C_\epsilon(x, \chi)$, describes the spatial correlation of the random field ϵ. Furthermore, the covariance function induces a linear operator, $\mathbf{T}_\epsilon \in \mathcal{L}(L^2(D), L^2(D))$, defined by

$$\mathbf{T}_\epsilon \phi(x) = \int_D C_\epsilon(x, \chi)\phi(\chi)\mathrm{d}\chi.$$

$\mathbf{T}_\epsilon$ is a compact, positive, and self adjoint operator. Furthermore, Mercer's Theorem ensures the existence of an orthonormal basis of eigenfunctions $\{\epsilon_k\}_{k=1}^\infty \subset L^2(D)$ and eigenvalues $\{\lambda_k\}_{k=1}^\infty \subset (0, \infty)$ such that

$$C_\epsilon(x, \chi) = \sum_{k=1}^\infty \lambda_k \epsilon_k(x)\epsilon_k(\chi)$$

(see Theorem 11 in Chapter 30, Section 5 of [69]). These eigenfunctions and eigenvalues induce the following decomposition of the random field ϵ,

$$\epsilon(\omega, x) = E[\epsilon(\cdot, x)] + \sum_{k=1}^\infty \sqrt{\lambda_k}\epsilon_k(x)Y_k(\omega) \tag{2.4.1}$$

where $Y_k(\omega) \in L^2_P(\Omega)$ satisfy

$$E[Y_k] = 0 \quad \text{and} \quad E[Y_j Y_k] = \delta_{jk} \quad \forall\, j, k = 1, 2, \ldots$$

The expansion (2.4.1) is known as the KL expansion of the random field ϵ. The convergence rate of the partial sums of the expansion (2.4.1) is completely dependent on the decay of the eigenvalues, λ_k, which depend on the covariance function (see [106]). Such decay induces anisotropy in the random field ϵ. Anisotropy here means that some directions, Γ_k, have a larger effect on the random field ϵ than others. As mentioned above, a common practice is to replace the random field ϵ with a truncated

KL expansion, i.e.

$$\epsilon(\omega, x) \leftarrow \epsilon_M(\omega, x) := E[\epsilon(\cdot, x)] + \sum_{k=1}^{M} \sqrt{\lambda_k}\epsilon_k(x)Y_k(\omega).$$

2.5 Tensor Product Function Spaces

The general formulation dictates the use of the Bochner space, $L^q_\rho(\Gamma; \mathcal{V})$, as the stochastic state spaces. Furthermore, in constructing a method for the solution of $\tilde{e}(u(y), z; y) = 0$ for all $y \in \Gamma$ and fixed $z \in \mathcal{Z}$, the stochastic collocation method seeks to interpolate $u = u(y) \in \mathcal{V}$ on a finite set of knots in the parameter space, Γ. Therefore, an application of stochastic collocation will require a finite number of solves of the state equation $\tilde{e}(u(y), z; y) = 0$. As is common in approximation theory, convergence of this interpolant is highly dependent on the regularity of $u = u(y)$ with respect to the parameter, $y \in \Gamma$. Assumption 2.2.2 guarantees that the solution u is a member of the Bochner space

$$u \in \mathcal{U} := L^q_\rho(\Gamma; \mathcal{V}) \quad \text{for some} \quad 1 \leq q \leq \infty.$$

Moreover, Assumption 2.2.4 implies $u = u(y)$ exists for every $y \in \Gamma$. Hence, it is often realistic to assume

$$u \in C^0_\rho(\Gamma; \mathcal{V}).$$

In the case of linear elliptic PDEs with uncertain inputs, certain regularity of these coefficients on the parameters, $y \in \Gamma$, ensures $u \in C^\infty_\rho(\Gamma; \mathcal{V})$. In fact, Assumption 3.2.1 implies this fact for general constraints. In this section, I will develop theory for three classes of tensor product spaces. First, I will discuss the construction of Bochner spaces. Secondly, I will present a few results concerning $C^0_\rho(\Gamma; \mathcal{V})$, and finally, I will provide some results for $C^0_\rho(\Gamma)$.

The goal of this Bochner space discussion is to relate $L^q_\rho(\Gamma; \mathcal{V})$ with the spaces

$L^q_\rho(\Gamma)$ and $\mathcal{V}$. First, define the tensor product space

$$L^q_\rho(\Gamma) \otimes \mathcal{V} := \left\{ \sum_{n=1}^{N} f_n v_n \ : \ \{f_n\} \subset L^q_\rho(\Gamma), \ \{v_n\} \subset \mathcal{V}, \text{ and } N \in \mathbb{N} \right\}.$$

Such sums are well defined since functions in $L^q_\rho(\Gamma)$ output into $\mathbb{R}$ and $\mathcal{V}$ is a real Banach space. It is clear from the definition of $L^q_\rho(\Gamma) \otimes \mathcal{V}$ that

$$L^q_\rho(\Gamma) \otimes \mathcal{V} \subset L^q_\rho(\Gamma; \mathcal{V}),$$

i.e. for $v = \sum_{n=1}^{N} f_n v_n \in L^q_\rho(\Gamma) \otimes \mathcal{V}$, it is easy to see that

$$\|v\|^q_{L^q_\rho(\Gamma;\mathcal{V})} \leq \sum_{n=1}^{N} \|v_n\|_{\mathcal{V}} \|f_n\|_{L^q_\rho(\Gamma)} < \infty.$$

With the appropriate choice of norm on $L^q_\rho(\Gamma) \otimes \mathcal{V}$ (c.f. the projective norm, see Chapter 2.3 in [102]), one can show that the completion of $L^q_\rho(\Gamma) \otimes \mathcal{V}$ is isometrically isomorphic to $L^q_\rho(\Gamma; \mathcal{V})$. This choice of norm is associated with the natural norm on the Bochner space via the relationship for $f \in L^q_\rho(\Gamma)$ and $v \in \mathcal{V}$

$$\|f \otimes v\|^q_\pi := \int_\Gamma \rho(y) \|f(y)v\|^q_{\mathcal{V}} \mathrm{d}y.$$

Here, the tensor product, $\otimes$ is defined in the standard way, i.e. $(f \otimes v) : y \mapsto f(y)v$. Note that this product is well defined since $\mathcal{V}$ is closed under scalar multiplication. Hence, it is sufficient to approximate the elements $u \in L^q_\rho(\Gamma; \mathcal{V})$ by elements in the tensor product space, $\widehat{u} \in L^q_\rho(\Gamma) \otimes \mathcal{V}$.

In the same manner, one can relate $C^0_\rho(\Gamma; \mathcal{V})$ to the spaces $C^0_\rho(\Gamma)$ and $\mathcal{V}$. Define the tensor product space

$$C^0_\rho(\Gamma) \otimes \mathcal{V} := \left\{ \sum_{n=1}^{N} f_n v_n \ : \ \{f_n\} \subset C^0_\rho(\Gamma), \ \{v_n\} \subset \mathcal{V}, \text{ and } N \in \mathbb{N} \right\}.$$

Again, such sums are well defined since functions in $C^0_\rho(\Gamma)$ real valued and $\mathcal{V}$ is a real Banach space. Furthermore, it is clear that

$$C^0_\rho(\Gamma) \otimes \mathcal{V} \subset C^0_\rho(\Gamma; \mathcal{V})$$

by similar arguments as used in the Bochner space discussion. Now, with the appropriate choice of norm on $C^0_\rho(\Gamma)\otimes\mathcal{V}$ (c.f. the injective norm, see Chapter 3.2 in [102]), one can show that the completion of $C^0_\rho(\Gamma)\otimes\mathcal{V}$ is isometrically isomorphic to the Banach space, $C^0_\rho(\Gamma;\mathcal{V})$. This choice of norm is associated with the natural norm on the space $C^0_\rho(\Gamma;\mathcal{V})$via the relationship for $f\in C^0_\rho(\Gamma)$ and $v\in\mathcal{V}$

$$\|f\otimes v\|_\varepsilon := \sup_{y\in\Gamma}\|\rho(y)f(y)v\|_{\mathcal{V}}.$$

The tensor product $f\otimes v$ is defined as in the Bochner space discussion. Therefore, it is again sufficient to approximate a function $u\in C^0_\rho(\Gamma;\mathcal{V})$ by a function in the tensor space, $\widehat{u}\in C^0_\rho(\Gamma)\otimes\mathcal{V}$.

Since $C^0_\rho(\Gamma;\mathcal{V})$ can be associated with the completion of $C^0_\rho(\Gamma)\otimes\mathcal{V}$, I will focus on approximation in $C^0_\rho(\Gamma)\otimes\mathcal{V}$. In order to do this, I would like to associate $C^0_\rho(\Gamma)$ with the one dimensional function spaces $C^0_{\rho_k}(\Gamma_k)$ for $k=1,\ldots,M$. In this case, I can build approximation operators for $C^0_\rho(\Gamma)$ based on one dimensional operators acting on $C^0_{\rho_k}(\Gamma_k)$. As before, define the tensor product space

$$C^0_{\rho_1}(\Gamma_1)\otimes\cdots\otimes C^0_{\rho_M}(\Gamma_M) := \left\{\sum_{n=1}^{N}\prod_{k=1}^{M} f_{k,n} \;:\; \{f_{k,n}\}\subset C^0_{\rho_k}(\Gamma_k) \text{ and } N\in\mathbb{N}\right\}.$$

Clearly, this definition implies

$$C^0_{\rho_1}(\Gamma_1)\otimes\cdots\otimes C^0_{\rho_M}(\Gamma_M)\subset C^0_\rho(\Gamma)$$

and using the appropriate norm (again, the injective norm [102]), one can show that the completion of $C^0_{\rho_1}(\Gamma_1)\otimes\cdots\otimes C^0_{\rho_M}(\Gamma_M)$ is isometrically isomorphic to $C^0_\rho(\Gamma)$. The norm in this case is defined by the relationship for $f_k\in C^0_{\rho_k}(\Gamma_k)$, $k=1,\ldots,M$,

$$\|f_1\otimes\cdots\otimes f_M\|_\varepsilon := \sup_{y\in\Gamma}|\rho_1(y_1)f_1(y_1)\cdot\ldots\cdot\rho_M(y_M)f_M(y_M)|.$$

Here, the notation $f_1\otimes\cdots\otimes f_M$ is associated with the mapping

$$(f_1\otimes\cdots\otimes f_M): y\mapsto f_1(y_1)\cdot\ldots\cdot f_M(y_M).$$

With this said, the numerical methods in this thesis will attempt to approximate functions in $C^0_\rho(\Gamma) \otimes \mathcal{V}$. Furthermore, to approximate the parametric dependence in $C^0_\rho(\Gamma)$, it will suffice to only work in $C^0_{\rho_1}(\Gamma_1) \otimes \cdots \otimes C^0_{\rho_M}(\Gamma_M)$.

Chapter 3

The Stochastic Collocation Method

Stochastic collocation is a non-intrusive method for solving the high dimensional parametric equation (2.2.3),

$$\tilde{e}(u(y), z; y) = 0 \quad \text{a.e. in } \Gamma$$

for fixed $z \in \mathcal{Z}$. Stochastic collocation is an interpolation based method and relies on the regularity of the solution $u \in \mathcal{U}$ to (2.2.3) with respect to $y \in \Gamma$ in order to achieve rapid convergence rates. In this chapter I will review the stochastic collocation discretization technique for the solution of the parametric equations. I will then extend these techniques to the case of the optimization problem (2.2.9)

$$\min_{z \in \mathcal{Z}} \ \widehat{J}(z) := \sigma(j(u(y; z), z))$$

where $u(z) = u \in \mathcal{U} = L^p_\rho(\Gamma; \mathcal{V})$ is the solution to (2.2.3). Furthermore, I will formulate the stochastic collocation method using the abstract approximation operator,

$$\mathcal{L}_Q : C^0_\rho(\Gamma) \to C_Q(\Gamma) \subset C^0_\rho(\Gamma).$$

The set $C_Q(\Gamma)$ is a finite dimensional subspace of $C^0_\rho(\Gamma)$ and the operator, $\mathcal{L}_Q$, only requires a finite number of function evaluations at the points in the set, $\mathcal{N}_Q \subset \Gamma$ with $|\mathcal{N}_Q| = Q$. Associated with $\mathcal{L}_Q$ is the quadrature operator, $E_Q := E \circ \mathcal{L}_Q : C^0_\rho(\Gamma) \to \mathbb{R}$.

The operators, $\mathcal{L}_Q$ and E_Q, are typically sparse grid or tensor product operators. The notion of sparse grids is developed in Chapter 4. Concluding this chapter, I will prove error bounds for stochastic collocation applied to optimization problems governed by parametric equations.

3.1 Collocation for Parametric Equations

Consider the parametric equation

$$\tilde{e}(u(y), z; y) = 0 \quad \text{a.e. in } \Gamma \tag{3.1.1}$$

where $u \in C^0(\Gamma; \mathcal{V})$ and $z \in \mathcal{Z}$. This equation typically represents a parametrized PDE or PDE with uncertain coefficients. As in Chapter 2, $\tilde{e}(y) : \mathcal{V} \times \mathcal{Z} \to \mathcal{W}$ for $y \in \Gamma$. The stochastic collocation method builds an approximate solution to (3.1.1) on a finite set of "collocation points." The approximation operator, $\mathcal{L}_Q$, eluded to above is a linear operator which depends on a finite number of function evaluations at points in $\mathcal{N}_Q = \{y_1, \ldots, y_Q\} \subset \Gamma$ with $|\mathcal{N}_Q| = Q$. Furthermore, I will assume that $\mathcal{L}_Q$ has the form

$$(\mathcal{L}_Q f)(y) = \sum_{k=1}^{Q} P_k(y) f(y_k) \quad \forall\, f \in C^0_\rho(\Gamma)$$

where $P_k \in C_Q(\Gamma)$ for $k = 1, \ldots, Q$. In Chapter 4, I will construct specific operators, $\mathcal{L}_Q$, as tensor product or sparse grid approximation operators. In the case that $\mathcal{L}_Q$ is a tensor product operator or a sparse grid operator built on nested one dimensional interpolation knots, $\mathcal{L}_Q$ is interpolatory, i.e. for any $f \in C^0_\rho(\Gamma)$

$$(\mathcal{L}_Q f)(y) = f(y) \quad \forall\, y \in \mathcal{N}_Q,$$

or equivalently, $P_k(y_j) = \delta_{kj}$ for $k, j = 1, \ldots, Q$. On the other hand, if $\mathcal{L}_Q$ is a sparse grid operator built on non-nested or weakly nested one dimensional interpolation knots, then $\mathcal{L}_Q$ is not interpolatory. These interpolation results are presented Chapter 4, Proposition 4.2.5. For the purposes of this thesis, the finite dimensional

approximation space, $C_Q(\Gamma)$, is assumed to be a polynomial space and in the context of tensor product and sparse grid operators, $C_Q(\Gamma)$ can be explicitly determined (see Proposition 4.2.4 in Chapter 4).

With this definition of $\mathcal{L}_Q$, the stochastic collocation solution to (3.1.1) is

$$u_Q(y) = (\mathcal{L}_Q)u(y) = \sum_{k=1}^{Q} P_k(y)u(y_k).$$

To compute u_Q, one needs to solve (3.1.1) for all $y \in \mathcal{N}_Q$, where $\mathcal{N}_Q$ denotes the Q interpolation knots associated with $\mathcal{L}_Q$. Thus, one must solve

$$\tilde{e}(u(y), z; y) = 0, \quad y \in \mathcal{N}_Q$$

in order to compute u_Q.

3.2 Regularity and Interpolation

The stochastic collocation method depends on the point-wise solution of (3.1.1); therefore, sufficient regularity of the solution to (2.2.4), $u \in \mathcal{U}$, with respect to the parameters, $y \in \Gamma$, is essential. First of all, $u \in \mathcal{U} \cap C^0_\rho(\Gamma; \mathcal{V})$ in order for point evaluations of u to be well defined. Moreover, one must ensure that the error committed through the approximation operator, $\mathcal{L}_Q$, decays as the number of collocation points increases. In the context of polynomial approximation and interpolation, the following assumption on $u \in \mathcal{U}$ is sufficient (c.f. see Chapter 7, Section 8 in [45]).

Assumption 3.2.1 (Analyticity of the State) *For fixed $y_j^* \in \Gamma_j$ with $j \neq k$, define*

$$u_k^*(y_k) := u(y_1^*, \ldots, y_{k-1}^*, y_k, y_{k+1}^*, \ldots, y_M^*).$$

The function, $u_k^* : \Gamma_k \to \mathcal{V}$, has an analytic extension on the open elliptic disc, $D_{r_k} \subset \mathbb{C}$ containing Γ_k. Specifically, if $\Gamma_k = [-1, 1]$, then $D_{r_k}(\Gamma_k)$ is the region

bounded by the ellipse

$$\mathcal{E}_{r_k} := \left\{ z = t + is \in \mathbb{C} \ : \ t = \frac{r_k + r_k^{-1}}{2} \cos\phi, \ s = \frac{r_k - r_k^{-1}}{2} \sin\phi, \ \phi \in [0, 2\pi) \right\}. \tag{3.2.1}$$

Ellipses in the complex plane are natural choices when analyzing interpolation and quadrature errors because the space $L^2(D_{r_k})$ is a reproducing kernel Hilbert space, therefore point evaluation is a continuous linear functional. For more information on reproducing kernel Hilbert spaces and quadrature error analysis see [112]. There are possibly many conditions on $e(u, z; y)$ for which Assumption 3.2.1 holds. I will now present one such set of assumptions. The assumptions made here are extensions of Assumption 1 in [68]. Assumption 1 in [68] are specific for the case where $e(u, z; y)$ is linear in u and z. Furthermore, similar assumptions were made for the case of truncated KL expansions in Lemma 3.2 of [9]. The proof of Theorem 3.2.2 below follows similar arguments as the proof of Lemma 3.2 in [9].

Theorem 3.2.2 *Suppose the Implicit Function Theorem holds for* (3.1.1) *and suppose* Γ *is bounded. Let* $z \in \mathcal{Z}$ *be fixed,* $u(z) = u \in C^0_\rho(\Gamma; \mathcal{V})$ *solve* (3.1.1)*, and define for fixed* $y_j^* \in \Gamma_j$*,* $j \neq k$*,*

$$\tilde{e}_k(u, z; y_k) := \tilde{e}(u_k^*(y_k), z; y_1^*, \ldots, y_{k-1}^*, y_k, y_{k+1}^*, \ldots, y_M^*).$$

If for each $k = 1, \ldots, M$*, there exists* $b < +\infty$ *and* $C = C(z) \geq \sup_{y_k \in \Gamma_k} \|u_k^*(y_k; z)\|_{\mathcal{V}}$ *such that*

$$\sup_{y_k \in \Gamma_k} \|(\partial_u \tilde{e}_k(u(z), z; y_k))^{-1} (\partial_{y_k}^n \tilde{e}_k(u(z), z; y_k))\| \leq C b^n \quad \forall n \in \mathbb{N}, \tag{3.2.2a}$$

$$\sup_{y_k \in \Gamma_k} \|(\partial_u \tilde{e}_k(u(z), z; y_k))^{-1} (\partial_u^n \tilde{e}_k(u(z), z; y_k))\| \leq n!\, b^n \quad \forall n \in \mathbb{N}, \tag{3.2.2b}$$

then $u = u(z)$ *satisfies Assumptions 3.2.1.*

Proof: By implicitly differentiating the equation $\tilde{e}(u, z; y) = 0$ with respect to y_k and applying the triangle inequality,

$$\begin{aligned}\left\|\partial_{y_k}^n u_k^*(y_k)\right\|_{\mathcal{V}} \leq & \left\|(\partial_u \tilde{e}_k(u(z), z; y_k))^{-1}(\partial_{y_k}^n \tilde{e}_k(u(z), z; y_k))\right\| \\ & + \sum_{k=1}^{n} \binom{n}{k} \left\|(\partial_u \tilde{e}_k(u(z), z; y_k))^{-1}(\partial_u^k \tilde{e}_k(u(z), z; y_k))\right\| \left\|\partial_{y_k}^{n-k} u_k^*(y_k)\right\|_{\mathcal{V}}\end{aligned}$$

Define $R_n := \frac{1}{n!}\left\|\partial_{y_k}^n u_k^*(y_k)\right\|_{\mathcal{V}}$, then $R_0 \leq C$ and by the bounds (3.2.2)

$$R_n \leq Cb^n + \sum_{m=1}^{n} b^m R_{n-m} \leq C(2b)^n.$$

Thus, for all $\gamma \in \mathbb{C}$ such that $|\gamma - y_k| \leq \tau < \frac{1}{2b}$,

$$u_k^*(\gamma) = \sum_{n=0}^{\infty} R_n(\gamma - y_k)^n \quad \text{and} \quad \|u_k^*(\gamma)\|_{\mathcal{V}} \leq \frac{C}{1 - 2b\tau} < +\infty. \tag{3.2.3}$$

The relations (3.2.3) hold for all $y_k \in \Gamma_k$; therefore, u_k^* has an analytic extension on the region

$$\Sigma_k(\tau) := \{\gamma \in \mathbb{C} \ : \ |\gamma - y_k| \leq \tau \ \forall \, y_k \in \Gamma_k\}$$

i.e. the union of all balls of radius τ with center y_k for all $y_k \in \Gamma_k$. This implies that u_k^* has an analytic extension on any ellipse contained in the set $\Sigma_k(\tau)$. □

Remark 3.2.3 *A similar result holds in the case that* Γ *is unbounded. To arrive at this result, one requires an auxiliary distribution which decays sufficiently fast. For more information on unbounded random variables and auxiliary distributions, see [9].*

Throughout this chapter, I will use an assumption on the error associated with the interpolation operator, $\mathcal{L}_Q$. I will use a general error bound and track this error through the optimization problem to the optimal controls.

Assumption 3.2.4 *Suppose* $f : \Gamma \to \mathbb{R}$ *has an analytic extension, then*

$$\|f - \mathcal{L}_Q f\|_{L_\rho^\infty(\Gamma)} \leq CQ^{-\nu}$$

where Q denotes the number of interpolation knots associated with $\mathcal{L}_Q$, $C = C(f) > 0$, and $\nu = \nu(f) > 0$.

A common form of the quantity $C = C(f)$ is $C = \widehat{C} \sup_{y \in \mathcal{D}} |f(y)|$ where $\mathcal{D} \subset \mathbb{C}$. Moreover, Assumption 3.2.4 and the tensor product structure of $L^\infty_\rho(\Gamma; \mathcal{V})$ (see Section 2.5) imply that the error committed by the stochastic collocation method is

$$\|u - \mathcal{L}_Q u\|_{L^\infty_\rho(\Gamma;\mathcal{V})} \leq C(u(z))Q^{-\nu}. \tag{3.2.4}$$

This error bound is proved in Theorems 3.8 and 3.13 of [85] for anisotropic Smolyak sparse grids built on Clenshaw-Curtis and Gaussian knots respectively. Similar error bounds are proved in Theorems 4.1 and 6.2 of [9] for tensor product and isotropic Smolyak sparse grids built on Gaussian knots. The error bounds presented in Theorems 4.1 and 6.2 of [9] and Theorem 3.13 of [85] are bounds on the $L^2_\rho(\Gamma; \mathcal{V})$ error. These error bounds are derived by first determining the bound (3.2.4). Since ρ is a probability distribution, the $L^2_\rho(\Gamma; \mathcal{V})$ error can then be bounded above by (3.2.4). In Chapter 4, I will prove the error bound (3.2.4) for specific operators, $\mathcal{L}_Q$.

3.3 Collocation for Optimization

In this section, I will present two possible stochastic collocation discretizations of the optimization problem

$$\min_{z \in \mathcal{Z}} \widehat{J}(z) := \sigma(j(u(y; z), z)) \tag{3.3.1}$$

where $u(y; z) = u(y) \in \mathcal{V}$ for $y \in \Gamma$ solves (3.1.1). The first approach is to approximate the solution of (3.1.1) using stochastic collocation and plug the discretized solution into the objective function. This is a common technique in PDE constrained optimization and is a "discretize then optimize" approach. The second approach is to approximate the map,

$$y \in \Gamma \mapsto j(u(y), z),$$

using the approximation operator, $\mathcal{L}_Q$. This also is a "discretize then optimize" approach, but is equivalent to an "optimize then discretize" approach.

To discretize (3.3.1) using the first method, one approximates the solution to (3.1.1) using stochastic collocation and the approximation operator, $\mathcal{L}_Q$, i.e.

$$u_Q(y;z) = (\mathcal{L}_Q u(z))(y) = \sum_{k=1}^{Q} P_k(y) u(y_k;z)$$

where $u(y_k) := u(y_k;z) \in \mathcal{V}$ for $k = 1,\dots,Q$ solves

$$\tilde{e}(u(y_k),z;y_k) = 0 \quad \text{for } k = 1,\dots,Q. \tag{3.3.2}$$

Then, plugging the approximate solution, $u_Q(y;z)$, into the objective function results in the following discretization of (3.3.1)

$$\min_{z\in\mathcal{Z}} \ \widehat{J}_Q(z) := \sigma(j(u_Q(y;z),z)) = \sigma\left(j\Big(\sum_{k=1}^{Q} P_k(y)u(y_k;z),z\Big)\right). \tag{3.3.3}$$

The adjoint calculus of Section 2.3 is easily extended to (3.3.3) yielding the derivative

$$\widehat{J}_Q'(z) = \sum_{k=1}^{Q} \tilde{e}_z(u(y_k;z),z;y_k)^*\lambda_k + E\Big[\nabla\sigma(j(u_Q(y;z),z))j_z(u_Q(y;z),z)\Big]$$

where $\lambda_k = \lambda_k(z) \in \mathcal{W}^*$ solves the adjoint equation

$$\tilde{e}_u(u(y_k z),z;y_k)^*\lambda_k + E\Big[P_k(y)\nabla\sigma(j(u_Q(y;z),z))j_u(u_Q(y;z),z)\Big] = 0 \tag{3.3.4}$$

for all $k = 1,\dots,Q$. Recalling the infinite dimensional parametrized adjoint equation, (2.3.2), from Section 2.3,

$$\tilde{e}_u(u,z;y)^*\lambda + \nabla\sigma(j(u,z))j_u(u,z) = 0 \quad \text{for a.e. } y \in \Gamma, \tag{3.3.5}$$

it is not clear that (3.3.4) corresponds to a discretization of (3.3.5), especially when sparse grid operators, $\mathcal{L}_Q$, are used. Furthermore, notice that the adjoint equation (3.3.4) requires the explicit knowledge of the polynomials P_k for $k = 1,\dots,Q$. For the class of approximation operators, $\mathcal{L}_Q$, considered in this thesis, the polynomials, P_k

for $k = 1, \ldots, Q$, are challenging to compute (see Chapter 4 for a detailed description of the operator $\mathcal{L}_Q$ used in my computations). Therefore, the discretized optimization problem, (3.3.3), will not be used.

Instead of approximating the solution of (3.1.1) as $u_Q = \mathcal{L}_Q u$ and substituting u_Q into the objective function, I will approximate the map

$$y \in \Gamma \mapsto j(u(y), z)$$

for any $u \in \mathcal{U}$ using $\mathcal{L}_Q$. This again is a "discretize then optimize" approach where instead of discretizing the constraint, I have discretized the objective function. To do this, I will replace σ with $\sigma_Q := \sigma \circ \mathcal{L}_Q$. This substitution yields the optimization problem

$$\min_{z \in \mathcal{Z}} \widehat{J}_Q(z) := \sigma_Q(j(u(y; z), z)) = \sigma\left(\sum_{k=1}^{Q} P_k(y) j(u(y_k; z), z) \right) \tag{3.3.6}$$

where $u(y_k; z) = u(y_k) \in \mathcal{V}$ solves

$$\tilde{e}(u(y_k), z; y_k) = 0 \quad \forall\, k = 1, \ldots, Q.$$

Furthermore, the derivative of $\widehat{J}_Q(z)$ is

$$\widehat{J}_Q'(z) = \sum_{k=1}^{Q} \vartheta_k \Big\{ \tilde{e}_z(u(y_k; z), z; y_k)^* \lambda_k + j_z(u(y_k; z), z) \Big\}$$

where $\vartheta_k = E\Big[\nabla\sigma\Big(\sum_{\ell=1}^{Q} P_\ell(y) j(u(y_\ell; z), z) \Big) P_k(y) \Big]$ and $\lambda_k(z) = \lambda_k \in \mathcal{W}^*$ solves

$$\tilde{e}_u(u(y_k; z), z; y_k)^* \lambda_k + j_u(u(y_k; z), z) = 0 \quad \forall\, k = 1, \ldots, Q.$$

Note that if $\sigma \equiv E$, then ϑ_k are the quadrature weights associated with the tensor product quadrature rule, E_Q, i.e. $\omega_k := \vartheta_k = E[P_k]$. In this case, the notions of "optimize then discretize" and "discretize then optimize" coincide. Using this discretization scheme, the discretized state and adjoint equations correspond to the stochastic collocation discretization of the true state and adjoint equations, (2.2.4)

and (2.3.2) respectively. Although this discretization scheme is consistent, there is one apparent pitfall of using this general tensor product discretization. The pitfall is that the cubature weights associated with E_Q may not all be positive. Thus, it is possible that $\widehat{J}_Q(z)$ is not convex even if $\widehat{J}(z)$ is.

3.4 Collocation Error Bounds for Optimization

In discussing the discretization of the optimization problem (3.3.1), it is crucial to understand error between a given discretization and the true (infinite dimensional) solution. In this section, I will prove an error bound for the discretized problem, (3.3.6) corresponding to the test problem presented in Section 2.1. I will first prove the error bound for the risk measure, $\sigma(Y) = E[Y]$, then I will generalize this result to other common risk measures.

3.4.1 Minimizing the Expected Value

The expected value risk measure results in the linear quadratic optimal control problem

$$\min_{z\in\mathcal{Z}} \widehat{J}(z) := \frac{1}{2}E\Big[\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\Big] + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 \tag{3.4.1}$$

where $u(y) = u(y;z) \in \mathcal{V}$ for all $y \in \Gamma$ solves

$$\mathbf{A}(y)u(y) + \mathbf{B}(y)z + \mathbf{b}(y) = 0 \quad \forall\, y \in \Gamma.$$

Recall here that $\mathcal{Z}$ and $\mathcal{H}$ are Hilbert spaces, and $\mathcal{V}$ and $\mathcal{W}$ are Banach spaces. Furthermore, $\mathbf{Q} \in \mathcal{L}(\mathcal{V},\mathcal{H})$ is an observation operator, $\bar{q} \in \mathcal{H}$ is the desired state, $\mathbf{A}(y) \in \mathcal{L}(\mathcal{V},\mathcal{W})$ for all $y \in \Gamma$ is the state operator, $\mathbf{B}(y) \in \mathcal{L}(\mathcal{Z},\mathcal{W})$ for all $y \in \Gamma$ is the control operator, and $\mathbf{b}(y) \in \mathcal{W}$ for all $y \in \Gamma$ is an inhomogeneity. Moreover, $\sigma = E : L^1_\rho(\Gamma) \to \mathbb{R}$. Therefore, the state space is $\mathcal{U} = L^2_\rho(\Gamma;\mathcal{V})$ and the Lagrange multiplier space is $\mathcal{Y}^* = L^2_\rho(\Gamma;\mathcal{V}) = \mathcal{U}$.

The error analysis presented here is centered around the particular form of the derivatives of $\widehat{J}(z)$ and $\widehat{J}_Q(z)$. As such, I will now recall the specific forms of these derivatives. Since $\sigma(Y) = E[Y]$ is Fréchet differentiable, $\widehat{J}(z)$ is also Fréchet differentiable, and, as seen in Section 2.3, $\widehat{J}(z)$ has a gradient

$$\nabla \widehat{J}(z) = \alpha z + w(z)$$

where $w = w(z) \in \mathcal{Z}$ solves

$$\mathbf{R}w = E[\mathbf{B}^*(y)\lambda(y;z)]$$

and $\lambda(y) = \lambda(y;z) \in \mathcal{W}^*$ for all $y \in \Gamma$ solves the adjoint equation

$$\mathbf{A}^*(y)\lambda(y) + \mathbf{Q}^*(\mathbf{Q}u(y;z) - \bar{q}) = 0 \quad \forall\, y \in \Gamma. \tag{3.4.2}$$

Moreover, employing the discretization in (3.3.6) results in the following discretized gradient

$$\nabla \widehat{J}_Q(z) = \alpha z + w_Q(z)$$

where $w_Q = w_Q(z) \in \mathcal{Z}$ solves

$$\mathbf{R}w_Q = E_Q[\mathbf{B}^*\lambda(z)] = \sum_{k=1}^{Q} \omega_k \mathbf{B}_k^* \lambda_k(z),$$

$\lambda_k = \lambda_k(z) \in \mathcal{W}^*$ solves the adjoint equation

$$\mathbf{A}_k^*\lambda_k + \mathbf{Q}^*(\mathbf{Q}u_k(z) - \bar{q}) = 0 \quad \forall\, k = 1, \ldots, Q,$$

and $u_k = u_k(z) \in \mathcal{V}$ solves the state equation

$$\mathbf{A}_k u_k + \mathbf{B}_k z + \mathbf{b}_k = 0 \quad \forall\, k = 1, \ldots, Q.$$

Here, note that since $u_k(z) = u(y_k;z)$ where $u(y;z)$ solves $\tilde{e}(u(y),z;y) = 0$, one also has that $\lambda_k(z) = \lambda(y_k;z)$ where $\lambda(y;z)$ solves the adjoint equation, (2.3.2). Now, since $\mathbf{R}$ is the linear operator representing the $\mathcal{Z}$ inner product, $\mathbf{R}$ is a positive, invertible operator, i.e. there exists a bounded inverse $\mathbf{R}^{-1} \in \mathcal{L}(\mathcal{Z}^*, \mathcal{Z})$, and $w = w(z) \in \mathcal{Z}$

can be written as $w(z) = \mathbf{R}^{-1}E[\mathbf{B}^*(y)\lambda(y;z)]$. Similarly, for the discretized problem, $w_Q(z) = \mathbf{R}^{-1}E_Q[\mathbf{B}^*\lambda(z)]$. Computing the difference between the true gradient and the discretized gradient, thus gives

$$\nabla\widehat{J}(z) - \nabla\widehat{J}_Q(z) = w(z) - w_Q(z) = \mathbf{R}^{-1}(E - E_Q)[\mathbf{B}^*\lambda(z)]. \tag{3.4.3}$$

This clearly shows that the error in the gradient vectors is controlled by the error associated with the quadrature operator, E_Q, or equivalently by the error associated with the interpolation operator, $\mathcal{L}_Q$.

Theorem 3.4.1 *Suppose $z^* \in \mathcal{Z}$ satisfies the first order necessary conditions for* (3.3.1) *and suppose $z_Q^* \in \mathcal{Z}$ satisfies the first order necessary conditions for* (3.3.6)*. Then, the error between z^* and z_Q^* satisfies*

$$\alpha\|z^* - z_Q^*\|_{\mathcal{Z}} \leq E\Big[\|\mathbf{R}^{-1}\mathbf{B}^*\lambda(z_Q^*) - \mathcal{L}_Q\mathbf{R}^{-1}\mathbf{B}^*\lambda(z_Q^*)\|_{\mathcal{Z}}\Big].$$

Proof: Since $\widehat{J}(z)$ is uniformly convex, it satisfies the inequality

$$\alpha\|z^* - z_Q^*\|_{\mathcal{Z}}^2 \leq \langle\nabla\widehat{J}(z^*) - \nabla\widehat{J}(z_Q^*), z^* - z_Q^*\rangle_{\mathcal{Z}},$$

and the Cauchy-Schwarz inequality implies

$$\alpha\|z^* - z_Q^*\|_{\mathcal{Z}} \leq \|\nabla\widehat{J}(z^*) - \nabla\widehat{J}(z_Q^*)\|_{\mathcal{Z}}.$$

Now since $z^* \in \mathcal{Z}$ is a first order critical point of $\widehat{J}(z^*)$ and $z_Q^* \in \mathcal{Z}$ is a first order critical point of $\widehat{J}_Q(z)$, i.e. $\nabla\widehat{J}(z^*) = 0$ and $\nabla\widehat{J}_Q(z_Q^*) = 0$, the above inequality can be rewritten as

$$\alpha\|z^* - z_Q^*\|_{\mathcal{Z}} \leq \|\nabla\widehat{J}_Q(z_Q^*) - \nabla\widehat{J}(z_Q^*)\|_{\mathcal{Z}}.$$

Plugging (3.4.3) into the right hand side of the above inequality gives

$$\alpha\|z^* - z_Q^*\|_{\mathcal{Z}} \leq \|\mathbf{R}^{-1}(E - E_Q)[\mathbf{B}^*\lambda(z_Q^*)]\|_{\mathcal{Z}},$$

and an application of Jensen's inequality yields the desired result. □

From this theorem, the following corollary follows directly from Assumption 3.2.4.

Corollary 3.4.2 *Suppose the solution to the adjoint equation, $\lambda \in \mathcal{U}$, satisfy Assumption 3.2.1 and $\mathcal{L}_Q$ satisfies Assumption 3.2.4. Then there exists positive constants $C = C(p(z_Q^*), \alpha)$ and ν such that*

$$\|z^* - z_Q^*\|_{\mathcal{Z}} \leq CQ^{-\nu}.$$

Proof: This is a consequence of Theorem 3.4.1 and the error bound in Assumption 3.2.4. Assumption 3.2.4 is applicable due to the tensor product nature of $\mathcal{Y}^* = \mathcal{U}$ (see Section 2.5 for more details). □

Remark 3.4.3 *Recall the linear quadratic control problem described in Section 2.1. For this model problem, $\mathbf{A}$ is defined by*

$$\langle \mathbf{A}(y)v, w\rangle_{\mathcal{V}^*,\mathcal{V}} = \int_D \epsilon(y,x)\nabla v(x)\cdot\nabla w(x)\mathrm{d}x \quad \forall\, v,\, w \in \mathcal{V}$$

and is uniformly bounded above for all $y \in \Gamma$ because $\epsilon \in L^\infty(\Gamma \times D)$. Furthermore, $\mathbf{A}(y)$ has an almost everywhere bounded inverse because of the uniform ellipticity assumption (2.1.11)*. On the other hand, $\mathbf{b}(y) \equiv 0$ and $\mathbf{B}$ is independent of $y \in \Gamma$. Therefore, if there exists $C > 0$ and $b < \infty$ such that*

$$\frac{\partial^n \epsilon}{\partial y_k^n}(y,x) \leq Cb^n$$

for almost all $(y,x) \in \Gamma \times D$ and for all $n \in \mathbb{N}$, then Theorem 3.2.2 ensures that the solution to the state equation, u, has an analytic extension. Moreover, these conditions and the analyticity of u guarantee that the solution to the adjoint equation, λ, also has an analytic extension. Thus, Corollary 3.4.2 applies to this test problem.

3.4.2 Minimizing the Mean Plus Semi-Deviation

The expected value problem (3.4.1) is typically not a sufficient reformulation of (2.2.9). In many engineering applications, a design or control holding "on average" is unacceptable. It is often necessary to account for tail probabilities and extreme events.

Such "robust" optimization problems are formulated using risk measures. I will now generalize the results of Theorem 3.4.1 and Corollary 3.4.2 to the mean plus semi-deviation coherent risk measure. Recall that, in general, the mean plus semi-deviation of order q is defined as

$$\sigma_q(Y) := E[Y] + cE\Big[[Y - E[Y]]_+^q\Big]^{1/q}$$

for $c \in [0,1]$ and $\sigma_q(Y)$ can be written in terms of the auxiliary function

$$\widehat{\sigma}_q : \mathbb{R} \times L_\rho^q(\Gamma) \to \mathbb{R}$$

as $\sigma_q(Y) = \widehat{\sigma}_Q(E[Y], Y)$ where

$$\widehat{\sigma}_q(t, Y) := t + cE\Big[[Y - t]_+^q\Big]^{1/q}.$$

This characterization will allow for easy derivative computation. Since $\sigma_q(Y)$ satisfies Assumptions 2.2.11 (i.e. $\sigma_q(Y)$ is coherent) and the deterministic objective function is separable, the reformulated objective function corresponding to mean plus semi-deviation can be written as

$$\widehat{J}(z) = \frac{1}{2}\sigma_1\Big(\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2.$$

This objective function, $\widehat{J}(z)$, is Hadamard differentiable and admits a gradient since $\mathcal{Z}$ is a Hilbert space. This Hadamard differentiability follows from the discussion on page 16 of [100]. Define the set

$$\mathcal{Y}(z) := \left\{ y \in \Gamma \ : \ \|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 > E\Big[\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\Big] \right\},$$

and let $\chi_{\mathcal{Y}(z)}$ denote the characteristic function of the set $\mathcal{Y}(z)$, then the gradient in $\mathcal{Z}$ corresponding to the Hadamard derivative is given by

$$\begin{aligned}\nabla\widehat{J}(z) &= \alpha z + \mathbf{R}^{-1}\left\{ E[\mathbf{B}^*\lambda] - cE[\chi_{\mathcal{Y}(z)}]E[\mathbf{B}^*\lambda] + cE[\chi_{\mathcal{Y}(z)}\mathbf{B}^*\lambda] \right\} \\ &= \alpha z + \mathbf{R}^{-1}\left\{ E[\mathbf{B}^*\lambda] + c\mathrm{Cov}\Big(\chi_{\mathcal{Y}(z)}, \mathbf{B}^*\lambda\Big) \right\} \end{aligned} \tag{3.4.4}$$

where $\lambda = \lambda(z)$ solves the adjoint equation (3.4.2). Here, $\mathrm{Cov}(Y_1, Y_2)$ denotes the covariance between the random variables Y_1 and Y_2. The covariance is defined as

$$\mathrm{Cov}(Y_1, Y_2) = E[(Y_1 - E[Y_1])(Y_2 - E[Y_2])] = E[Y_1 Y_2] - E[Y_1]E[Y_2].$$

To discretize this objective function, $\widehat{J}(z)$, replace $\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2$ with its interpolant, i.e.

$$\widehat{J}_Q(z) = \frac{1}{2}\sigma_1\Big(\mathcal{L}_Q\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$$

where $\mathcal{L}_Q$ is an appropriate interpolation operator. The linearity of $\mathcal{L}_Q$ and the gradient (3.4.4) implies the following form of the gradient of $\widehat{J}_Q(z)$

$$\nabla\widehat{J}_Q(z) = \alpha z + \mathbf{R}^{-1}\left\{E[\mathcal{L}_Q\mathbf{B}^*\lambda] + c\mathrm{Cov}\Big(\chi_{\mathcal{Y}_Q(z)}, \mathcal{L}_Q\mathbf{B}^*\lambda\Big)\right\} \tag{3.4.5}$$

where

$$\mathcal{Y}_Q(z) := \left\{y \in \Gamma \ : \ \mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 > E\Big[\mathcal{L}_Q\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\Big]\right\}.$$

As was done in the proof of Theorem 3.4.1, I will need to quantify the error between $\nabla\widehat{J}(z)$ and $\nabla\widehat{J}_Q(z)$ at the same control value. To do this, first notice that the covariance operator satisfies

$$\mathrm{Cov}(Y_1, Y_2) - \mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(Y_1, Y_2 - X_2) + \mathrm{Cov}(Y_1 - X_1, X_2)$$

for any random variables $Y_1, Y_2, X_1, X_2 \in L_\rho^2(\Gamma)$. Applying Hölder's inequality to this equality gives

$$\begin{aligned}|\mathrm{Cov}(Y_1, Y_2) - \mathrm{Cov}(X_1, X_2)| \leq &\|Y_1 - E[Y_1]\|_{L_\rho^\infty(\Gamma)}\|(Y_2 - X_2) - E[Y_2 - X_2]\|_{L_\rho^1(\Gamma)} \\ &+ \|X_2 - E[X_2]\|_{L_\rho^\infty(\Gamma)}\|(Y_1 - X_1) - E[Y_1 - X_1]\|_{L_\rho^1(\Gamma)}.\end{aligned} \tag{3.4.6}$$

Now notice that an application of the triangle inequality to $\|Y - E[Y]\|$ for any $L_\rho^q(\Gamma)$ norm, $\|\cdot\|$, gives

$$\|Y - E[Y]\| \leq \|Y\| + \|E[Y]\| = \|Y\| + |E[Y]| \leq 2\|Y\|.$$

Therefore, multiple applications of the triangle inequality to the right hand side of (3.4.6) gives

$$|\mathrm{Cov}(Y_1,Y_2)-\mathrm{Cov}(X_1,X_2)| \le 4\Big\{ \|Y_1\|_{L^\infty_\rho(\Gamma)}\|Y_2-X_2\|_{L^1_\rho(\Gamma)}+\|X_2\|_{L^\infty_\rho(\Gamma)}\|Y_1-X_1\|_{L^1_\rho(\Gamma)}\Big\}.$$

Plugging in $Y_1=\chi_{\mathcal{Y}(x)}$, $Y_2=\mathbf{B}^*p$, $X_1=\chi_{\mathcal{Y}_Q(x)}$, and $X_2=\mathcal{L}_Q\mathbf{B}^*p$ gives the following error representation for the gradients

$$\begin{aligned}\|\nabla\widehat{J}(z)-\nabla\widehat{J}_Q(z)\|_{\mathcal{Z}} \le & \underbrace{E\Big[\|\mathbf{R}^{-1}\mathbf{B}^*\lambda(z)-\mathcal{L}_Q\mathbf{R}^{-1}\mathbf{B}^*\lambda(z)\|_{\mathcal{Z}}\Big]}_{\text{I}} \\ & + 4\underbrace{\Big\|\sup_{y\in\Gamma}|\mathbf{R}^{-1}\mathbf{B}(y)^*\lambda(y;z)-\mathbf{R}^{-1}\mathcal{L}_Q\mathbf{B}(y)^*\lambda(y;z)|\Big\|_{\mathcal{Z}}}_{\text{II}} \\ & + 4\underbrace{\Big\|\sup_{y\in\Gamma}|\mathbf{R}^{-1}\mathcal{L}_Q\mathbf{B}(y)^*\lambda(y;z)|\Big\|_{\mathcal{Z}}E[\chi_{\mathcal{Y}(z)\Delta\mathcal{Y}_Q(z)}]}_{\text{III}}\end{aligned}$$

where $\mathcal{Y}(z)\Delta\mathcal{Y}_Q(z)=(\mathcal{Y}(z)\backslash\mathcal{Y}_Q(z))\cup(\mathcal{Y}_Q(z)\backslash\mathcal{Y}(z))$ denotes the symmetric difference of the sets $\mathcal{Y}(z)$ and $\mathcal{Y}_Q(z)$. Notice that expressions I and II can be handled using existing error bounds for the interpolation operator $\mathcal{L}_Q$ (see Theorem 3.4.1), therefore it remains to bound III. To do this, one must bound the size (i.e. probability) of the symmetric difference $\mathcal{Y}(z)\Delta\mathcal{Y}_Q(z)$.

In general, determining meaningful bounds on the size of $\mathcal{Y}(z)\Delta\mathcal{Y}_Q(z)$ is not possible. To circumvent this issue, it is convenient to replace $[\,\cdot\,]_+$ with a C^∞ approximation. A common choice of smooth approximation is

$$\wp(x,\gamma)=x+\frac{1}{\gamma}\log(1+e^{-\gamma x})$$

(c.f. see [38] for more details). This function has many nice properties; most importantly, $\wp(x,\gamma)$ tends to $[x]_+$ as γ grows to positive infinity. Other significant properties are

$$\wp(x,\gamma)>[x]_+, \quad |\wp'(x,\gamma)|<1, \quad \text{and} \quad |\wp''(x,\gamma)|\le\frac{\gamma}{4} \quad \forall\, x\in\mathbb{R}.$$

Furthermore, $\wp(x,\gamma)$ is strictly convex and strictly increasing (see Lemma 1.1 in [38]). The fact that $\wp''(x,\gamma)$ is bounded for all $x\in\mathbb{R}$ implies that, for fixed $\gamma\in(0,\infty)$,

$\wp'(x,\gamma)$ is Lipschitz continuous. Now, define the C^∞ approximation of $\sigma_q(Y)$ as

$$\sigma_q^\gamma(Y) := E[Y] + cE\Big[\wp(Y - E[Y],\gamma)^q\Big]^{1/q}$$

for fixed γ. The above analysis can be replicated by replacing σ_q with σ_q^γ, which gives the following gradients

$$\begin{aligned}\nabla\widehat{J}(z) =& \alpha z + \mathbf{R}^{-1}E[\mathbf{B}^*p] \\ &+ c\mathbf{R}^{-1}\mathrm{Cov}\left(\wp'\Big(\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - E\Big[\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\Big],\gamma\Big), \mathbf{B}^*\lambda\right) \end{aligned} \tag{3.4.7}$$

and

$$\begin{aligned}\nabla\widehat{J}_Q(z) =& \alpha z + \mathbf{R}^{-1}E[\mathcal{L}_Q\mathbf{B}^*p] \\ &+ c\mathbf{R}^{-1}\mathrm{Cov}\left(\wp'\Big(\mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - E\Big[\mathcal{L}_Q\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\Big],\gamma\Big), \mathcal{L}_Q\mathbf{B}^*\lambda\right). \end{aligned} \tag{3.4.8}$$

Now, computing the difference between these gradients and employing the properties of $\wp(x,\gamma)$ described above gives the bound

$$\begin{aligned}\|\nabla\widehat{J}(z) - \nabla\widehat{J}_Q(z)\|_{\mathcal{Z}} \leq \mathrm{I} + 4\times\mathrm{II} + 2\gamma\Big\|\sup_{y\in\Gamma}\big|\mathbf{R}^{-1}\mathcal{L}_Q\mathbf{B}(y)^*\lambda(y;z)\big|\Big\|_{\mathcal{Z}} \\ \times E\Big[\big|\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - \mathcal{L}_Q\|\mathbf{Q}u(z) - \bar{q}\|_{\mathcal{H}}^2\big|\Big]. \end{aligned} \tag{3.4.9}$$

This bound demonstrates the error in the gradients is controlled by the interpolation error associated with $\mathcal{L}_Q$. Hence, for this C^∞ approximation to the mean plus semi-deviation problem, one gets a similar error bound on the gradients as with the expected value problem.

Theorem 3.4.4 *Suppose $z^* \in \mathcal{Z}$ is a first order critical point of*

$$\widehat{J}(z) = \frac{1}{2}\sigma_1^\gamma\Big(\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$$

and $z_Q^* \in \mathcal{Z}$ is a first order critical point of

$$\widehat{J}_Q(z) = \frac{1}{2}\sigma_1^\gamma\Big(\mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2.$$

Furthermore, suppose the assumptions of Corollary 3.4.2 hold. Then there exists positive constants $C = C(u(z_Q^*), p(z_Q^*), \gamma, \alpha)$ *and* ν *such that*

$$\|z^* - z_Q^*\|_{\mathcal{Z}} \leq CQ^{-\nu}.$$

Proof: $\widehat{J}(z)$ is uniformly convex since $\sigma_1^\gamma(Y)$ is convex and increasing. Therefore, the error in the controls can be bounded by the errors in the gradients, $\nabla\widehat{J}(z_Q^*)$ and $\nabla\widehat{J}_Q(z_Q^*)$ as done in the proof of Theorem 3.4.1. Using the bound (3.4.9) and the Assumption 3.2.4 gives the desired result. □

3.4.3 Minimizing the Conditional Value-At-Risk

It is often necessary to account for tail probabilities and extreme events. One method of formulating these "robust" optimization problems is to employ the conditional value-at-risk (CVaR). CVaR quantifies the risk on the tails of the distribution of the objective function and is defined as

$$\sigma_{\text{CVaR}}(Y) = \min_{t\in\mathbb{R}} t + cE\Big[[Y-t]_+\Big]$$

and much like the mean plus semi-deviation, CVaR can be written using the auxiliary function $\widehat{\sigma}_q(t, Y)$ as $\sigma_{\text{CVaR}}(Y) = \min_{t\in\mathbb{R}} \widehat{\sigma}_1(t, Y)$. It can further be shown that minimizing the CVaR objective function over $\mathcal{Z}$ is equivalent to minimizing $\widehat{\sigma}_1(t, Y)$ over the augmented space $(t, z) \in \mathbb{R} \times \mathcal{Z}$ [117]. For this analysis, I will consider the optimization problem

$$\min_{t\in\mathbb{R},\, z\in\mathcal{Z}} \widehat{J}(t, z) = \widehat{\sigma}_1\Big(t, \frac{1}{2}\|\mathbf{Q}u(y; z) - \bar{q}\|_{\mathcal{H}}^2\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$$

where $u(y; z) = u \in \mathcal{U}$ solves (2.2.4). The analysis performed for the mean plus semi-deviation risk measure essential holds for the CVaR risk measure and therefore it will be critical to replace $[\,\cdot\,]_+$ in σ_{CVaR} with $\wp(\cdot, \gamma)$. The C^∞ approximation of the CVaR objective function is

$$\widehat{J}(t, z) = \widehat{\sigma}_1^\gamma\Big(t, \frac{1}{2}\|\mathbf{Q}u(y; z) - \bar{q}\|_{\mathcal{H}}^2\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 \tag{3.4.10}$$

where

$$\widehat{\sigma}_1^{\gamma}(Y,t) = t + cE\Big[\wp(Y-t,\gamma)\Big].$$

The gradient of $\widehat{J}(t,z)$ is thus computed using the partial derivatives of $\widehat{J}(t,z)$ with respect to $t \in \mathbb{R}$ and $z \in \mathcal{Z}$:

$$\begin{aligned}\nabla_t \widehat{J}(t,z) =& 1 - cE\Big[\wp'\big(\frac{1}{2}\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - t, \gamma\big)\Big] \\ \nabla_z \widehat{J}(t,z) =& \alpha z + c\mathbf{R}^{-1}E\Big[\wp'\big(\frac{1}{2}\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - t, \gamma\big)\mathbf{B}^*\lambda\Big]\end{aligned}$$

where $\lambda = \lambda(z)$ solves the adjoint equation (3.4.2).

Employing the stochastic collocation discretization scheme, one can write the collocation and C^{∞} approximate CVaR objective function as

$$\widehat{J}_Q(t,z) = \widehat{\sigma}_1^{\gamma}\Big(t, \frac{1}{2}\mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2\Big) + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2 \tag{3.4.11}$$

which admits the gradient

$$\begin{aligned}\nabla_t \widehat{J}_Q(t,z) =& 1 - cE\Big[\wp'\big(\frac{1}{2}\mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - t, \gamma\big)\Big] \\ \nabla_z \widehat{J}_Q(t,z) =& \alpha z + c\mathbf{R}^{-1}E\Big[\wp'\big(\frac{1}{2}\mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - t, \gamma\big)\mathcal{L}_Q\mathbf{B}^*\lambda\Big].\end{aligned}$$

Invoking the Lipschitz continuity of $\wp'(\cdot,\gamma)$ and Jensen's inequality, the collocation error associated with t component of the gradient is bounded by

$$\begin{aligned}|\nabla_t \widehat{J}(t,z) - \nabla_t \widehat{J}_Q(t,z)| \leq& cE\Big[\big|\wp'\big(\frac{1}{2}\mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - t, \gamma\big) \\ & - \wp'\big(\frac{1}{2}\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - t, \gamma\big)\big|\Big] \\ \leq& \frac{c\gamma}{8}E\Big[\big|\mathcal{L}_Q\|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2 - \|\mathbf{Q}u(y;z) - \bar{q}\|_{\mathcal{H}}^2\big|\Big].\end{aligned}$$

Therefore, the collocation error in the t component of the gradient of $\widehat{J}(t,z)$ is controlled by interpolation error. Similarly, the collocation error associated with the z

component of the gradient is bounded above in the following manner

$$\begin{aligned}
\|\nabla_z \widehat{J}(t,z) - \nabla_z \widehat{J}_Q(t,z)\|_{\mathcal{Z}} =& c\Big\|\mathbf{R}^{-1}E\Big[\wp'\big(\frac{1}{2}\|\mathbf{Q}u(y;z)-\bar{q}\|^2_{\mathcal{H}} - t,\gamma\big)\mathbf{B}^*\lambda \\
&- \wp'\big(\frac{1}{2}\mathcal{L}_Q\|\mathbf{Q}u(y;z)-\bar{q}\|^2_{\mathcal{H}} - t,\gamma\big)\mathcal{L}_Q\mathbf{B}^*\lambda\Big]\Big\|_{\mathcal{Z}} \\
\leq& cE\Big[\big|\wp'\big(\frac{1}{2}\|\mathbf{Q}u(y;z)-\bar{q}\|^2_{\mathcal{H}} - t,\gamma\big)\big|\|\mathbf{R}^{-1}(\mathbf{B}^*\lambda - \mathcal{L}_Q\mathbf{B}^*\lambda)\|_{\mathcal{Z}} \\
&+ \big|\wp'\big(\frac{1}{2}\|\mathbf{Q}u(y;z)-\bar{q}\|^2_{\mathcal{H}} - t,\gamma\big) \\
&- \wp'\big(\frac{1}{2}\mathcal{L}_Q\|\mathbf{Q}u(y;z)-\bar{q}\|^2_{\mathcal{H}} - t,\gamma\big)\big|\|\mathbf{R}^{-1}\mathcal{L}_Q\mathbf{B}^*\lambda\|_{\mathcal{Z}}\Big] \\
\leq& cE\Big[\big|\|\mathbf{R}^{-1}(\mathbf{B}^*\lambda - \mathcal{L}_Q\mathbf{B}^*\lambda)\|_{\mathcal{Z}}\Big] \\
&+ \frac{c\gamma \sup_{y\in\Gamma}\|\mathbf{R}^{-1}\mathcal{L}_Q\mathbf{B}^*(y)\lambda(y)\|_{\mathcal{Z}}}{8} \qquad (3.4.12) \\
&\times E\Big[\big|\|\mathbf{Q}u(y;z)-\bar{q}\|^2_{\mathcal{H}} - \mathcal{L}_Q\|\mathbf{Q}u(y;z)-\bar{q}\|^2_{\mathcal{H}}\big|\Big]. \qquad (3.4.13)
\end{aligned}$$

Clearly, the error associated with the z component of the gradient can also be controlled by interpolation error. This brings about the following error bound.

Theorem 3.4.5 *Suppose* $(t^*, z^*) \in \mathbb{R}\times\mathcal{Z}$ *is a first order critical point of* (3.4.10) *and* $(t^*_Q, z^*_Q) \in \mathbb{R}\times\mathcal{Z}$ *is a first order critical point of* (3.4.11)*. Furthermore, suppose the assumptions of Corollary 3.4.2 hold. Then there exists positive constants* $C = C(u(z^*_Q), \lambda(z^*_Q), \gamma, \alpha)$ *and* ν *such that*

$$\|z^* - z^*_Q\|_{\mathcal{Z}} \leq C(Q^{-\nu} + |t^* - t^*_Q|).$$

Proof: $\widehat{J}(t,z)$ is uniformly convex with respect to $z\in\mathcal{Z}$ and strictly convex with respect to $t\in\mathbb{R}$ since $\widehat{\sigma}^\gamma_1(t,Y)$ is convex and increasing in both arguments. Therefore, the error in the controls can be bounded by the errors in the gradients, $\nabla\widehat{J}(t^*_Q, z^*_Q)$ and $\nabla\widehat{J}_Q(t^*_Q, z^*_Q)$ as done in the proof of Theorem 3.4.1, i.e.

$$\begin{aligned}
\alpha\|z^* - z^*_Q\|_{\mathcal{Z}} \leq& \|\nabla_z\widehat{J}(t^*,z^*) - \nabla_z\widehat{J}(t^*,z^*_Q)\|_{\mathcal{Z}} \\
=& \|\nabla_z\widehat{J}_Q(t^*_Q,z^*_Q) - \nabla_z\widehat{J}(t^*,z^*_Q)\|_{\mathcal{Z}} \\
\leq& \|\nabla_z\widehat{J}_Q(t^*_Q,z^*_Q) - \nabla_z\widehat{J}(t^*_Q,z^*_Q)\|_{\mathcal{Z}} + \|\nabla_z\widehat{J}(t^*_Q,z^*_Q) - \nabla_z\widehat{J}(t^*,z^*_Q)\|_{\mathcal{Z}}.
\end{aligned}$$

Applying the bound (3.4.13) and Assumption 3.2.4 to the first term on the right hand side and invoking the Lipschitz continuity of $\wp'$ to handle the second term one gets

$$\alpha\|z^* - z_Q^*\|_{\mathcal{Z}} \leq \widehat{C}(u(z_Q^*), \lambda(z_Q^*), \gamma)Q^{-\nu} + \frac{c\gamma}{4}E\Big[\|\mathbf{R}^{-1}\mathbf{B}^*\lambda(z_Q^*)\|_{\mathcal{Z}}\Big]|t^* - t_Q^*|.$$

This gives the desired result. □

Remark 3.4.6 *To obtain a meaningful error bound, one also needs to quantify the error, $|t^* - t_Q^*|$, in terms of Q. Currently, this is future work. Note that the techniques used to bound the error $\|z^* - z_Q^*\|_{\mathcal{Z}}$ cannot be applied to the error $|t^* - t_Q^*|$ since $\widehat{J}(t, z)$ is only strictly convex with respect to $t \in \mathbb{R}$.*

Chapter 4

High Dimensional Interpolation

In this chapter I will review and extend a general class of high dimensional interpolation and quadrature operators. These operators are essential for efficient application of the stochastic collocation discretization technique for the solution of the parametric equation $\tilde{e}(u(y), z; y) = 0$ for $y \in \Gamma$ when Γ is a high dimensional parameter space. I will first review a few standard results concerning one dimensional Lagrangian interpolation. Then, I will extend the one dimensional interpolation operators to a general class of high dimensional interpolation operators built on the concepts of Smolyak's Algorithm [110] and weighted tensor approximation [120, 52, 15]. Finally, I will present an extension of the dimension adaptive ideas of Gerstner and Griebel [52] for the approximation of high dimensional integrals.

4.1 One Dimensional Interpolation

The one dimensional interpolation operators considered in this section will be denoted $\mathcal{L}_{k,j} : C^0_{\rho_k}(\Gamma_k) \to \mathbb{P}^{m_{k,j}-1}(\Gamma_k)$ for $k = 1, \ldots, M$. Here, $\mathcal{L}_{k,j}$, $j = 1, 2, \ldots$, denotes a sequence of one dimensional interpolation operators exact for polynomials of degree $m_{k,j} - 1$ or less and $\{m_{k,j}\}_{j=1}^{\infty}$ is a monotonically increasing sequence of positive integers with $m_{k,1} = 1$. Moreover, $\mathcal{N}_{k,j} = \{y_{k,j,q}\}_{q=1}^{m_{k,j}} \subset \Gamma_k$ will denote the finite set of

distinct interpolation knots corresponding to the operator $\mathcal{L}_{k,j}$. For any $f \in C^0_{\rho_k}(\Gamma_k)$, the one dimensional interpolation operators considered here are defined as

$$(\mathcal{L}_{k,j} f)(y) = \sum_{q=1}^{m_{k,j}} \ell_{k,j,q}(y) f(y_{k,j,q}).$$

The functions $\ell_{k,j,q} \in \mathbb{P}^{m_{k,j}-1}(\Gamma_k)$ are taken to be the Lagrange interpolating polynomials built on the one dimensional knots, $\mathcal{N}_{k,j}$, i.e.

$$\ell_{k,j,q}(y_k) = \prod_{\substack{s=1 \\ s \neq q}}^{m_{k,j}} \frac{y_k - y_{k,j,s}}{y_{k,j,q} - y_{k,j,s}}.$$

Other one dimensional interpolating polynomials such as piecewise polynomials or splines are also applicable to the tensor product constructions which follow, but I will restrict my attention to Lagrange interpolation.

The norm of the Lagrange interpolation operator, $\mathcal{L}_{k,j}$, is known as the *Lebesgue constant*, i.e.

$$\Lambda^p_{k,j} := \|\mathcal{L}_{k,j}\|_{k,p} \quad \text{for} \quad 1 \leq p \leq \infty,$$

where $\|\cdot\|_{k,p}$ denotes the operator norm

$$\|\mathcal{L}_{k,j}\|_{k,p} := \sup_{\|f\|_{L^p_{\rho_k}(\Gamma_k)} \leq 1} \|\mathcal{L}_{k,j} f\|_{L^p_{\rho_k}(\Gamma_k)}.$$

For the remainder of this thesis, I will restrict my attention to the case where $p = \infty$. The associated Lebesgue constant satisfies

$$\Lambda_{k,j} := \Lambda^\infty_{k,j} = \max_{y \in \Gamma_k} \sum_{q=1}^{m_{k,j}} |\ell_{k,j,q}(y)|.$$

Since $\Lambda_{k,j}$ is dependent on the choice of interpolation knots, it is valid to ask "what is the optimal choice of knots, $\mathcal{N}^*_{k,j}$," and "how does the Lebesgue constant for these optimal knots, $\Lambda^*_{k,j}$, behave." It can be shown that the Lebesgue constant, $\Lambda^*_{k,j}$, satisfies the lower bound

$$\Lambda_{k,j} \geq \Lambda^*_{k,j} > \frac{2}{\pi} \log(m_{k,j} - 1) + \frac{2}{\pi}\Big(\gamma + \log \frac{4}{\pi}\Big)$$

where $\gamma \approx 0.5772$ is Euler's constant (c.f. see the remark on page 701 of [30]). A consequence of this result is that $\Lambda_{k,j}$ grows unboundedly as j increases toward infinity.

A simple consequence of these definitions is any interpolation operator, $\mathcal{L}_{k,j}$ gives the following interpolation error for all $f \in C^0_{\rho_k}(\Gamma_k)$

$$\begin{aligned}\|f - \mathcal{L}_{k,j} f\|_{C^0_{\rho_k}(\Gamma_k)} &\leq \|f - p + \mathcal{L}_{k,j}(p - f)\|_{C^0_{\rho_k}(\Gamma_k)} \\ &\leq \|f - p\|_{C^0_{\rho_k}(\Gamma_k)} + \|\mathcal{L}_{k,j}(p - f)\|_{C^0_{\rho_k}(\Gamma_k)}\end{aligned}$$

for any $p \in \mathbb{P}^{m_{k,j}-1}(\Gamma_k)$. Therefore,

$$\|f - \mathcal{L}_{k,j} f\|_{C^0_{\rho_k}(\Gamma_k)} \leq (1 + \Lambda_{k,j}) \inf_{p \in \mathbb{P}_{m_{k,j}-1}(\Gamma_k)} \|f - p\|_{C^0_{\rho_k}(\Gamma_k)}. \tag{4.1.1}$$

Thus, to fully characterize the interpolation error, one must be able to bound the error between $f \in C^0_{\rho_k}(\Gamma_k)$ and its best approximation in $\mathbb{P}^{m_{k,j}-1}(\Gamma_k)$. On the other hand, the error (4.1.1) is highly dependent on the size of the Lebesgue constant, $\Lambda_{k,j}$, which depends on the choice of interpolation knots, $\mathcal{N}_{k,j}$. In addition to (4.1.1), the error corresponding to the interpolation operator, $\mathcal{L}_{k,j}$, satisfies

$$\|\mathbf{I}_k - \mathcal{L}_{k,j}\|_{k,\infty} = 1 + \Lambda_{k,j} \tag{4.1.2}$$

where $\mathbf{I}_k$ denotes the identity operator on $C^0_{\rho_k}(\Gamma_k)$ (c.f. see Equation 8 and the proof that follows in [97]). This also gives the upper and lower bounds on the difference between two consecutive interpolation rules, $\Delta_{k,j} := \mathcal{L}_{k,j} - \mathcal{L}_{k,j-1}$,

$$|\Lambda_{k,j} - \Lambda_{k,j-1}| \leq \|\Delta_{k,j}\|_{k,\infty} \leq \Lambda_{k,j} + \Lambda_{k,j-1}.$$

Another property of these knots, $\mathcal{N}_{k,j}$, that will be essential for the high dimensional generalizations of these one dimensional interpolation operators is whether or not these knots are nested.

Definition 4.1.1 *The one dimensional interpolation abscissa are nested if*

$$\mathcal{N}_{k,j} \subset \mathcal{N}_{k,j+1} \quad \forall j.$$

They are weakly nested if

$$\mathcal{N}_{k,j} \cap \mathcal{N}_{k,j+1} \neq \emptyset, \quad but \quad \mathcal{N}_{k,j} \not\subset \mathcal{N}_{k,j+1} \quad \forall j.$$

See table 4.1 for common choices of one dimensional abscissa. Nested knots and to

Abscissa	Domain	Weight	m_i	Nested?
Clenshaw-Curtis	$[-1,1]$	$\rho(y)=1$	$2^{i-1}+1$	Yes
Gauss-Patterson	$[-1,1]$	$\rho(y)=1$	$2^{i+1}-1$	Yes
Gauss-Legendre	$[-1,1]$	$\rho(y)=1$	$2^{i+1}-1$	Weakly
Gauss-Hermite	$(-\infty,+\infty)$	$\rho(y)=e^{-y^2}$	$2^{i+1}-1$	Weakly
Genz-Keister	$(-\infty,+\infty)$	$\rho(y)=e^{-y^2}$	$\{1,3,9,19,35,...\}$	Yes
Gauss-Legendre	$[0,+\infty)$	$\rho(y)=e^{-y}$	$2^{i+1}-1$	No

Table 4.1: Common one dimensional abscissa with exponential growth rules, m_i.

some extent, weakly nested knots will give severe reductions in the number of high dimensional interpolation knots required to achieve a desired accuracy.

4.1.1 Interpolation and Analytic Functions

The Stone-Weierstrass Theorem implies that the space of polynomials is dense in $C^0_{\rho_k}(\Gamma_k)$ (c.f. see Theorem 4.45 in [49]); that is, any function, $f \in C^0_{\rho_k}(\Gamma_k)$ can be approximated arbitrarily closely by a polynomial. This fact does not ensure that a given sequence of interpolation operators converges for all members of $C^0_{\rho_k}(\Gamma_k)$. In fact, for each $\mathcal{L}_{k,j}$ there exists a function $f \in C^0_{\rho_k}(\Gamma_k)$ for which $(\mathcal{L}_{k,j}f)(y)$ diverges almost everywhere (e.g. Runge's function for equi-distant interpolation knots). This result is presented as Theorem 5.3 in [45]. This is no longer the case when the domain of the operator $\mathcal{L}_{k,j}$ is switched from $f \in C^0_{\rho_k}(\Gamma_k)$ to the analytic functions defined on an elliptic disc in the complex plane. This function space will be denoted by

$$\mathcal{A}(\Gamma_k, r_k) := \{f : \mathbb{C} \to \mathbb{R} \ : \ f \text{ is analytic on } D_{r_k}(\Gamma_k)\}.$$

The set, $D_{r_k}(\Gamma_k)$, is an open elliptic disc containing Γ_k with semi-axis sum $r_k > 1$ (e.g. see (3.2.1) for an example of $D_{r_k}(\Gamma_k)$ with $\Gamma_k = [-1, 1]$). The analyticity assumption, Assumption 3.2.1, ensures that for fixed directions, Γ_k, the solution of the PDE with uncertain coefficients (2.2.4) is a member of $\mathcal{A}(\Gamma_k, r_k)$. One result of particular interest to this work is Theorem 8.1 in Chapter 7.8 of [45]. This theorem concerns the best approximation of an analytic function, $f \in \mathcal{A}(\Gamma_k, r_k)$ with algebraic polynomials. A consequence of the proof of this theorem is the following lemma.

Lemma 4.1.2 *Suppose $f \in C^0_{\rho_k}(\Gamma_k)$ has an analytic extension on the elliptic disc, $D_{r_k}(\Gamma_k)$. Then, the best approximation of f by the algebraic polynomials of degree d, $\mathbb{P}_d(\Gamma_k)$ satisfies*

$$\min_{p \in \mathbb{P}_d(\Gamma_k)} \|f - p\|_{C^0_{\rho_k}(\Gamma_k)} \leq \frac{2}{r_k - 1} r_k^{-d} \sup_{y \in D_{r_k}} |f(y)|.$$

Combining this result with (4.1.1), it is easy to see that for any $f \in C^0_{\rho_k}(\Gamma_k)$ which has an analytic extension on the elliptic disc, $D_{r_k}(\Gamma_k)$, the interpolation error associated with the operator, $\mathcal{L}_{k,j}$, satisfies

$$\|f - \mathcal{L}_{k,j} f\|_{C^0_{\rho_k}(\Gamma_k)} \leq (1 + \Lambda_{k,j}) \frac{2}{r_k - 1} r_k^{-(m_{k,j}-1)} \sup_{y \in D_{r_k}} |f(y)|. \tag{4.1.3}$$

The convergence rate described in (4.1.3) depends on the sum of the semi-axis lengths, $r_k > 1$, of the ellipse, $D_{r_k}(\Gamma_k)$ which is determined by the analyticity of the function $f \in C^0_{\rho_k}(\Gamma_k)$.

4.1.2 Clenshaw-Curtis Knots

One popular choice of interpolation knots are the so-called Clenshaw-Curtis knots. These knots are the extrema of Chebyshev polynomials, i.e.

$$y_{k,j,q} = -\cos\left(\frac{\pi(q-1)}{m_{k,j} - 1}\right) \quad \text{for} \quad q = 1, \ldots, m_{k,i}.$$

Furthermore, these knots can be made nested by choosing $m_{k,i}$ to satisfy

$$m_{k,1} = 1 \quad \text{and} \quad m_{k,j} = 2^{j-1} + 1 \quad \text{for} \quad j > 1,$$

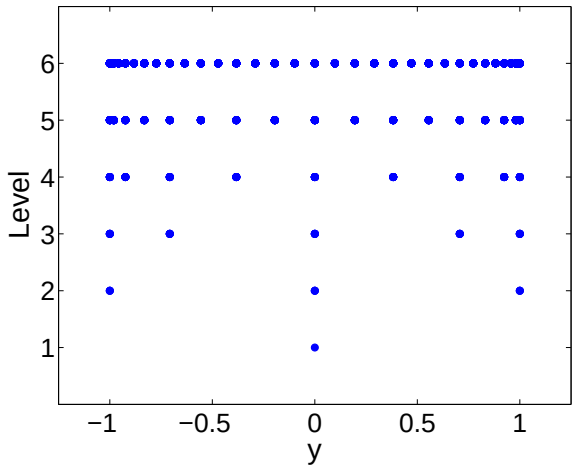


Figure 4.1: Clenshaw-Curtis quadrature nodes for levels $j = 1, \ldots, 6$.

see Figure 4.1. These knots are quite popular because they admit a relatively small upper bound for the Lebesgue constant. This upper bound is

$$\Lambda_{k,j} \leq \frac{2}{\pi} \log(m_{k,j} - 1) + 1 - \alpha_j < \frac{2}{\pi} \log(m_{k,j} - 1) + 1$$

for $m_{k,j} \geq 2$ and some $\alpha_j \in (0, 1/m_{k,j})$ (see e.g. [30]). In addition, plugging the particular form of $m_{k,j}$ into the bound for the Lebesgue constant gives

$$\Lambda_{k,j} \leq \frac{2}{\pi}(j-1)\log(2) + 1 = \frac{2}{\pi}\log(2)j + (1 - \frac{2}{\pi}\log(2)).$$

Combining this result with (4.1.3) the interpolation error for functions $f \in C^0_{\rho_k}(\Gamma_k)$ with analytic extension on $D_{r_k}(\Gamma_k)$ satisfies

$$\|f - \mathcal{L}_{k,j} f\|_{C^0_{\rho_k}(\Gamma_k)} \leq C(r_k, f) j r_k^{-2^j} \tag{4.1.4}$$

where $C = C(r_k, f)$ is a positive constant depending on r_k and f, but not on the polynomial degree, j. This bound on the one dimensional interpolation error also gives a bound on the difference of two consecutive operators, $\Delta_{k,j} = \mathcal{L}_{k,j} - \mathcal{L}_{k,j-1}$,

$$\|\Delta_{k,j} f\|_{C^0_{\rho_k}(\Gamma_k)} \leq \|\mathcal{L}_{k,j} f - f\|_{C^0_{\rho_k}(\Gamma_k)} + \|f - \mathcal{L}_{k,j-1} f\|_{C^0_{\rho_k}(\Gamma_k)} \leq D(r_k, f) j r_k^{-2^{j-1}} \tag{4.1.5}$$

where $D = D(r_k, f)$ is a positive constant depending on r_k and f, but not on the polynomial degree, j. These difference operators will be paramount in the construction of general high dimensional interpolation operators.

4.2 Tensor Product Polynomial Approximation

The high dimensional interpolation operators considered in this thesis are natural extensions of the one dimensional operators discussed in Section 4.1. These high dimensional operators are built on the differences between two consecutive interpolation operators, i.e.

$$\Delta_{k,j} = \mathcal{L}_{k,j} - \mathcal{L}_{k,j-1} \quad \text{where} \quad \mathcal{L}_{k,0} = 0$$

for $j \geq 1$ and $k = 1, \ldots, M$. For all functions, $f \in C^0_{\rho_k}(\Gamma_k)$, with an analytic extension on $D_{r_k}(\Gamma_k)$, the difference, $\Delta_{k,j} f$, converges to zero. Furthermore, for $k = 1, \ldots, M$, these difference operators satisfy the telescoping sum properties:

$$\Delta_{k,1} = \mathcal{L}_{k,1} \quad \text{and} \quad \mathcal{L}_{k,n} = \sum_{j=1}^{n} \Delta_{k,j}.$$

The tensor product interpolation operators considered here are built on high dimensional tensor products of these one dimensional difference operators. These tensor product difference operators are defined as

$$\Delta_{\mathbf{j}} = \Delta_{1,j_1} \otimes \cdots \otimes \Delta_{M,j_M}$$

for multi-indices $\mathbf{j} = (j_1, \ldots, j_M) \in \mathbb{N}^M$. The M dimensional generalization of the one dimensional operator $\mathcal{L}_{k,j}$ is then given by

$$\mathcal{L}_{\mathcal{I}} := \sum_{\mathbf{i} \in \mathcal{I}} \Delta_{\mathbf{i}} \tag{4.2.1}$$

for some index set, $\mathcal{I} \subset \mathbb{N}^M$. The one goal in selecting an appropriate index set, $\mathcal{I}$, is to maintain the telescoping sum property from the one dimensional case. This

property will guarantee certain polynomial exactness and interpolation results concerning $\mathcal{L}_{\mathcal{I}}$ by ensuring that similar properties of the one dimensional operators hold. I thoroughly discuss these properties in Subsection 4.2.3. These results will require some notion of admissibility for index sets, $\mathcal{I} \subset \mathbb{N}^M$. For clarity, I will adopt the following notation and partial ordering on $\mathbb{N}^M$, for $\mathbf{i}, \mathbf{j} \in \mathbb{N}^M$,

$$\mathbf{j} \leq \mathbf{i} \iff j_k \leq i_k \quad \forall k = 1, \ldots, M.$$

Definition 4.2.1 *The set $\mathcal{I} \subset \mathbb{N}^M$ is admissible if $\mathbf{i} \in \mathcal{I}$ and $\mathbf{j} \leq \mathbf{i}$, then $\mathbf{j} \in \mathcal{I}$.*

Remark 4.2.2 *The classic Smolyak and full tensor product operators are special cases of* (4.2.1)*. The Smolyak rule of level $\ell > 0$ corresponds to the index set*

$$\mathcal{I}_{SM}(\ell, M) := \left\{ \mathbf{i} \in \mathbb{N}^M \ : \ \sum_{k=1}^{M} (i_k - 1) \leq \ell \right\},$$

while the full tensor product algorithm of level ℓ corresponds to the index set

$$\mathcal{I}_{TP}(\ell, M) := \left\{ \mathbf{i} \in \mathbb{N}^M \ : \ \max_{k=1,\ldots,M} (i_k - 1) \leq \ell \right\}.$$

See figure 4.2 for a depiction of $\mathcal{I}_{SM}(7, 2)$ and $\mathcal{I}_{TP}(7, 2)$. On the other hand, given any mapping $g : \mathbb{N}^M \to \mathbb{N}$, strictly increasing in each argument, the anisotropic tensor product operator of order ℓ is defined as

$$\mathcal{I}(\ell, M, g) := \left\{ \mathbf{i} \in \mathbb{N}^M \ : \ g(\mathbf{i}) \leq \ell \right\}.$$

Notice that $\mathcal{I}$ subsumes both the classic Smolyak index set, $\mathcal{I}_{SM}(\ell, M)$, and the full tensor product index set, $\mathcal{I}_{TP}(\ell, M)$; that is, the classic Smolyak index set and full tensor product index are defined with

$$g(\mathbf{i}) = \sum_{k=1}^{M} (i_k - 1) \quad \textit{and} \quad g(\mathbf{i}) = \max_{k=1,\ldots,M} (i_k - 1),$$

respectively.

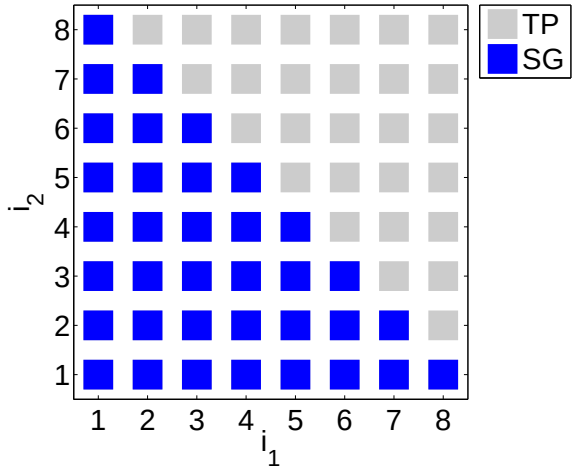


Figure 4.2: Tensor product (blue and grey) and Smolyak sparse grid (blue) index sets for level $\ell = 7$ and dimension $M = 2$.

The polynomial properties of the operators, $\mathcal{L}_\mathcal{I}$ are often disguised by the complexity in the definition (4.2.1). Hence, it is often useful to rewrite (4.2.1) as a linear combination of the tensor product of one dimensional operators, $\mathcal{L}_\mathbf{i} := \mathcal{L}_{1,i_1} \otimes \cdots \otimes \mathcal{L}_{M,i_M}$. This reformulation is called the combination technique [52, 50]. Let

$$\chi_\mathcal{I}(\mathbf{i}) = \begin{cases} 1 & \text{if } \mathbf{i} \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

denote the characteristic function of the index set $\mathcal{I}$. By expanding the M dimensional difference operators $\Delta_\mathbf{i}$ in terms of $\mathcal{L}_{k,j}$ and combining like terms, one can rewrite (4.2.1) as

$$\mathcal{L}_\mathcal{I} = \sum_{\mathbf{i} \in \mathcal{I}} \left(\sum_{\mathbf{z} \in \{0,1\}^M} (-1)^{|\mathbf{z}|} \chi_\mathcal{I}(\mathbf{i} + \mathbf{z}) \right) \mathcal{L}_\mathbf{i}. \tag{4.2.2}$$

This form of $\mathcal{L}_\mathcal{I}$ is convenient because it demonstrates the dependence of $\mathcal{L}_\mathcal{I}$ on the one dimensional interpolation operators. Furthermore, (4.2.2) demonstrates the savings achieved by (4.2.1) for a given choice of $\mathcal{I}$. Notice that if $\mathbf{i} \in \mathcal{I}$ such that $\mathbf{i} + \mathbf{z} \in \mathcal{I}$

for all $\mathbf{z} \in \{0,1\}^M$, then

$$c(\mathbf{i}) := \sum_{\mathbf{z}\in\{0,1\}^M} (-1)^{|\mathbf{z}|} \chi_{\mathcal{I}}(\mathbf{i}+\mathbf{z}) = \sum_{k=0}^{M} (-1)^k \begin{pmatrix} M \\ k \end{pmatrix} = 0.$$

Therefore, for such $\mathbf{i} \in \mathcal{I}$, one need not compute $\mathcal{L}_{\mathbf{i}}$.

It is worthwhile noting that much of the work associated with the operator $\mathcal{L}_{\mathcal{I}}$ is directly related to the index set, $\mathcal{I}$. This index set should be chosen so that $\mathcal{L}_{\mathcal{I}}$ is sufficiently accurate, but is not overly expensive to compute. To do this, many researchers have considered using *a priori* information to determine an "optimal" index set $\mathcal{I}$. Optimality, in this sense, refers to minimizing the error $\sum_{\mathbf{i}\notin\mathcal{I}} \|\Delta_{\mathbf{i}}\|$ subject to a constraint that the cost associated with $\mathcal{I}$ is less than some positive constant N. This constrained minimization problem is equivalent to a classic knapsack problem [13]. $\mathcal{I}$ can also be chosen in an adaptive fashion. In Subsection 4.2.2, I will discuss the adaptive selection of index sets, $\mathcal{I}$, in the context of sparse grid quadrature.

4.2.1 Tensor Product Quadrature

The tensor product operator, $\mathcal{L}_{\mathcal{I}}$, can be extended to a high dimensional quadrature. Let $E = E_1 \otimes \cdots \otimes E_M$ denote the M dimensional integral

$$E[f] = \int_{\Gamma} \rho(y) f(y) \mathrm{d}y = \prod_{k=1}^{M} \int_{\Gamma_k} \rho_k(y_k) f_k(y_k) \mathrm{d}y_k \quad \forall\, f \in C^0_{\rho_1}(\Gamma_1) \otimes \cdots \otimes C^0_{\rho_M}(\Gamma_M).$$

Associated with the tensor product operator, $\mathcal{L}_{\mathcal{I}}$, is the M dimensional cubature operator defined by the composition $E_{\mathcal{I}} := E \circ \mathcal{L}_{\mathcal{I}}$. The operator, $E_{\mathcal{I}}$, is the sparse grid cubature formula associated with the index set, $\mathcal{I}$. The combination technique formulation, (4.2.2), of $E_{\mathcal{I}}$ is

$$E_{\mathcal{I}} = E \circ \mathcal{L}_{\mathcal{I}} = \sum_{\mathbf{i}\in\mathcal{I}} \left(\sum_{\mathbf{z}\in\{0,1\}^M} (-1)^{|\mathbf{z}|} \chi_{\mathcal{I}}(\mathbf{i}+\mathbf{z}) \right) E_{\mathbf{i}}.$$

Here, $E_{\mathbf{i}} := E \circ \mathcal{L}_{\mathbf{i}}$ is merely the tensor product cubature rule associated with the index, $\mathbf{i}$. The cubature formula, $E_{\mathbf{i}}$, requires function evaluations at the nodes

$$\mathcal{N}_{\mathbf{i}} := \mathcal{N}_{1,i_1} \times \cdots \times \mathcal{N}_{M,i_M}.$$

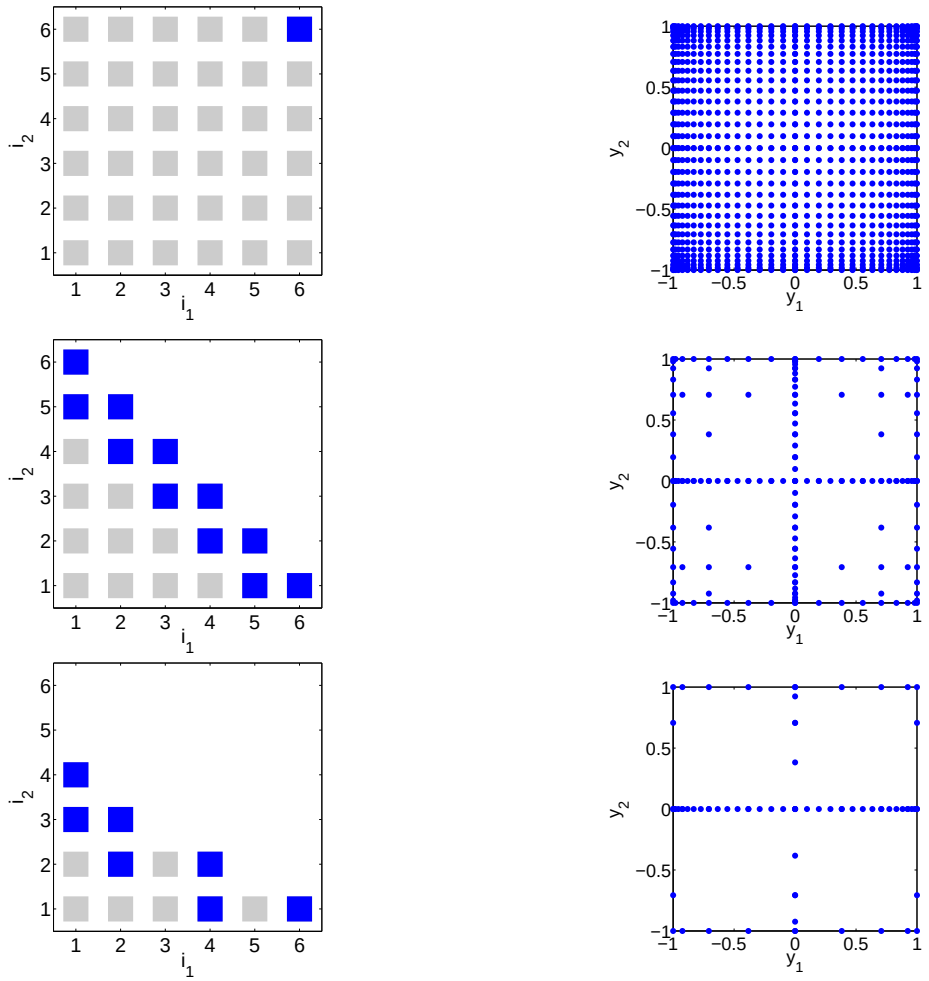


Figure 4.3: Admissible index sets (left) and their corresponding Clenshaw-Curtis quadrature nodes (right). The first is a tensor product rule, the second row is an isotropic Smolyak rule, and the third row is an arbitrary anisotropic rule. The blue and grey squares indicate members of the index sets. The blue squares correspond to indices for which $c(\mathbf{i}) = \sum_{\mathbf{z}\in\{0,1\}^M} (-1)^{|\mathbf{z}|} \chi_{\mathcal{I}}(\mathbf{i}+\mathbf{z}) \neq 0$.

Recall that the interpolation operator, $\mathcal{L}_{\mathbf{i}}$, is the tensor product of one dimensional Lagrange interpolation operators, $\mathcal{L}_{i_k}$, built on the nodes, $\mathcal{N}_{k,i_k}$. The weights associated with $E_{\mathbf{i}}$ are the Kronecker products of the corresponding one dimensional weights. In the case that $\mathcal{L}_{k,j}$ are one dimensional Lagrange interpolants built on the extrema of orthogonal polynomials, the cubature rules $E_{\mathbf{i}}$ are merely tensor product

Gaussian cubature rules. On the other hand, if $\mathcal{L}_{k,j}$ are one dimensional piecewise linear polynomials interpolants built on equidistant points, then $E_{\mathbf{i}}$ denotes a tensor product trapezoidal rule. Figure 4.3 depicts examples of the quadrature points associated with $E_{\mathcal{I}}$ for different choices of index sets, $\mathcal{I}$.

4.2.2 Dimension Adaptive Index Set Selection

Gerstner and Griebel [53], and Hegland [58] have presented an algorithm for the adaptive selection of admissible index sets, $\mathcal{I}$, when approximating the integral of a scalar valued function. This work can be extended to the general tensor product integration problem

$$\int_{\Gamma} \rho(y) f(y) \mathrm{d}y \quad \text{for } f \in C^0_\rho(\Gamma; \mathcal{V})$$

where $\mathcal{V}$ is some Banach space. In the context of optimization problems governed by PDEs with random inputs, such integration problems arise in gradient and Hessian-times-a-vector computations. The extended adaptive tensor product algorithm requires a reduction function, $\gamma : \mathcal{V} \to [0, \infty)$. One possible choice of γ is a norm defined on $\mathcal{V}$. See Algorithm (4.2.3) for details. This algorithm generates admissible index sets and the corresponding general tensor product cubature formula. Convergence of this algorithm is dependent on the quality of the error estimators, $\gamma(\Delta_{\mathbf{j}} v)$.

Algorithm 4.2.3 - Dimension Adaptive Sparse Grids:

Set $\mathbf{i} = (1, \ldots, 1)$, $\mathcal{O} = \emptyset$, $\mathcal{A} = \{\mathbf{i}\}$, $r = \Delta_{\mathbf{i}} v$, $\eta = \eta_{\mathbf{i}} = \gamma(r)$

while $\eta > TOL$ ***do***

Select $\mathbf{i} \in \mathcal{A}$ *corresponding to the largest* $\eta_{\mathbf{i}}$

Set $\mathcal{A} \leftarrow \mathcal{A} \setminus \{\mathbf{i}\}$ *and* $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathbf{i}\}$

Update the error indicator $\eta = \eta - \eta_{\mathbf{i}}$

for $k = 1, \ldots, d$ ***do***

Set $\mathbf{j} = \mathbf{i} + \mathbf{e}_k$

if $\mathcal{O} \cup \{\mathbf{j}\}$ *is admissible* ***then***

Set $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathbf{j}\}$

Set $\tilde{r} = \Delta_{\mathbf{j}} v$

Set $\eta_{\mathbf{j}} = \gamma(\tilde{r})$

Update the integral approximation $r = r + \tilde{r}$

Update the error indicator $\eta = \eta + \eta_{\mathbf{j}}$

end if

end for

end while

The convergence of Algorithm 4.2.3 is dependent on the regularity of the integrand, $f \in C^0_\rho(\Gamma; \mathcal{V})$ and is heuristic in nature. In practice, Algorithm 4.2.3 seems to work quite well and typically results in a large reduction of the number of function evaluations required to compute the integral. I will use Algorithm 4.2.3 as a means to adaptively approximate the optimization problems of interest. I will discuss the application of Algorithm 4.2.3 in Chapter 5.

4.2.3 Properties of the Tensor Product Operator, $\mathcal{L}_\mathcal{I}$

I will now discuss some consequences of the tensor product operators defined in (4.2.1). Namely, I will prove a result concerning the interpolation properties of $\mathcal{L}_\mathcal{I}$. In general, $\mathcal{L}_\mathcal{I}$ need not be interpolatory. In this subsection, I will prove a result characterizing in which cases $\mathcal{L}_\mathcal{I}$ is, in fact, interpolatory. Furthermore, I will discuss the polynomial exactness properties of $\mathcal{L}_\mathcal{I}$.

Associated with each operator, $\mathcal{L}_\mathcal{I}$ is a finite set of interpolation knots often called a "sparse grid." The combination technique, (4.2.2), gives an efficient representation

of this sparse grid through the coefficients

$$c(\mathbf{i}) := \sum_{\mathbf{z}\in\{0,1\}^M} (-1)^{|\mathbf{z}|}\chi_{\mathcal{I}}(\mathbf{i}+\mathbf{z}) \neq 0.$$

As discussed above, the only indices of consequence to the operator, $\mathcal{L}_{\mathcal{I}}$, are those for which $c(\mathbf{i}) \neq 0$, i.e. $\widehat{\mathcal{I}} := \{\mathbf{i} \in \mathcal{I} \ : \ c(\mathbf{i}) \neq 0\}$. Each tensor product operator $\mathcal{L}_{\mathbf{i}}$ for $\mathbf{i} \in \widehat{\mathcal{I}}$ requires function evaluations at tensor products of knots from the one dimensional interpolation abscissa, $\mathcal{N}_{k,j}$. The set of tensor product knots (sparse grid) associated with $\mathcal{L}_{\mathcal{I}}$ is thus defined as

$$\mathcal{N}_{\mathcal{I}} := \bigcup_{\mathbf{i}\in\widehat{\mathcal{I}}} (\mathcal{N}_{1,i_1} \times \cdots \times \mathcal{N}_{M,i_M}). \tag{4.2.3}$$

Notice that the size the sparse grid is bounded above by

$$|\mathcal{N}_{\mathcal{I}}| \leq \sum_{\mathbf{i}\in\widehat{\mathcal{I}}} |\mathcal{N}_{1,i_1}| \cdot \ldots \cdot |\mathcal{N}_{M,i_M}| = \sum_{\mathbf{i}\in\widehat{\mathcal{I}}} \prod_{k=1}^{M} m_{k,i_k}$$

and further reduction of the size of $\mathcal{N}_{\mathcal{I}}$ is achieved if the one dimensional knots, $\mathcal{N}_{k,j}$, are nested or weakly nested.

4.2.3.1 Polynomial Exactness

To fully describe the polynomial space for which $\mathcal{L}_{\mathcal{I}}$ is exact, notice that for $k = 1, \ldots, M$, the sequence $\{m_{k,j}\}_{j=1}^{\infty}$ is associated with functions $m_k : \mathbb{N} \to \mathbb{N}$ by the relation, $m_k(i) = m_{k,i}$. This mapping is monotone and injective. Therefore, m_k has a left inverse. In fact, one possible left inverse of m_k is

$$m_k^{-1}(p) = \min\left\{i \in \mathbb{N} \ : \ m_k(i) \geq p\right\}.$$

This choice of m_k^{-1} is an increasing function and satisfies the following two properties

$$m_k^{-1}(m_k(i)) = i \quad \text{and} \quad m_k(m_k^{-1}(p)) \geq p$$

see [12, 9] for more details. To simplify notation, let

$$\mathbf{m}(\mathbf{i}) := (m_1(i_1), \ldots, m_M(i_M)) \quad \text{and} \quad \mathbf{m}^{-1}(\mathbf{p}) := (m_1^{-1}(p_1), \ldots, m_M^{-1}(p_M)).$$

Now, define the set of polynomial degrees

$$\Lambda(\mathcal{I},\mathbf{m}) := \left\{\mathbf{p}\in\mathbb{N}^M \ : \ \mathbf{m}^{-1}(\mathbf{p}+1)\in\mathcal{I}\right\}.$$

The following proposition shows that the tensor product operator, $\mathcal{L}_{\mathcal{I}}$, is exact for all polynomials with degree in $\Lambda(\mathcal{I},\mathbf{m})$.

Proposition 4.2.4 *Suppose $\mathcal{I}\subset\mathbb{N}^M$ is admissible and the corresponding operator, $\mathcal{L}_{\mathcal{I}}$, defined by* (4.2.1) *is built on one dimensional Lagrange interpolating polynomials. Then for any $f\in C^0_\rho(\Gamma)$,*

$$\mathcal{L}_{\mathcal{I}} f\in\mathbb{P}(\Gamma,\mathcal{I},\mathbf{m}) := span\left\{\prod_{k=1}^M y_k^{p_k} \ : \ \mathbf{p}\in\Lambda(\mathcal{I},\mathbf{m}),\ y\in\Gamma\right\}. \tag{4.2.4}$$

Furthermore, $\mathcal{L}_{\mathcal{I}}\, p = p$ for any $p\in\mathbb{P}_{\mathcal{I}}$.

Proof: Define $\mathbb{P}_{\mathbf{j}}(\Gamma) := \text{span}\left\{\prod_{k=1}^M y_k^{p_k} \ : \ \mathbf{p}\le\mathbf{j},\ y\in\Gamma\right\}$ and notice that for any $f\in C^0(\Gamma)$, one has $\Delta_{\mathbf{i}} f\in\mathbb{P}_{\mathbf{m(i)}-1}(\Gamma)$. Therefore,

$$\begin{aligned}
\mathcal{L}_{\mathcal{I}} f &\in \text{span}\left\{\bigcup_{\mathbf{i}\in\mathcal{I}}\mathbb{P}_{\mathbf{m(i)}-1}(\Gamma)\right\}\\
&= \text{span}\left\{\bigcup_{\mathbf{i}\in\mathcal{I}}\text{span}\left\{\prod_{k=1}^M y_k^{p_k} \ : \ \mathbf{p}\le\mathbf{m(i)}-1\right\}\right\}\\
&= \text{span}\left\{\bigcup_{\mathbf{i}\in\mathcal{I}}\text{span}\left\{\prod_{k=1}^M y_k^{p_k} \ : \ \mathbf{m}^{-1}(\mathbf{p}+1)\le\mathbf{i}\right\}\right\}\\
&= \text{span}\left\{\prod_{k=1}^M y_k^{p_k} \ : \ \mathbf{m}^{-1}(\mathbf{p}+1)\in\mathcal{I}\right\}\\
&= \mathbb{P}_{\mathcal{I}}(\Gamma).
\end{aligned}$$

Note that the third equality follows since $\mathcal{I}$ is admissible. This proves (4.2.4).

Now, I will prove that $\mathcal{L}_{\mathcal{I}}$ is exact for any $f\in\mathbb{P}(\Gamma,\mathcal{I},\mathbf{m})$. By linearity of $\mathcal{L}_{\mathcal{I}}$, one only needs to prove that $\mathcal{L}_{\mathcal{I}}$ is exact for the monomial, $f(y)=\prod_{k=1}^M y_k^{p_k}$, with $\mathbf{p}\in\Lambda(\mathcal{I},\mathbf{m})$. Fix $\mathbf{p}\in\Lambda(\mathcal{I},\mathbf{m})$, then for any $\mathbf{i}\in\mathcal{I}$,

$$(\Delta_{\mathbf{i}} f)(y) = \Delta_{\mathbf{i}}\prod_{k=1}^M y_k^{p_k} = \prod_{k=1}^M(\mathcal{L}_{k,i_k}-\mathcal{L}_{k,i_k-1})y_k^{p_k}.$$

Since $\mathcal{L}_{k,j}$ is exact for polynomials of degree $m_k(j)-1$ or less, $(\mathcal{L}_{k,i_k}-\mathcal{L}_{k,i_k-1})y_k^{p_k}=0$ whenever $p_k \leq m_k(i_k-1)-1$. Now, define the index

$$\bar{\mathbf{i}} := (\bar{i}_1,\ldots,\bar{i}_M) \quad \text{with} \quad \bar{i}_k = m_k^{-1}(p_k+1), \quad k=1,\ldots,M.$$

The index $\bar{\mathbf{i}}$ satisfies $\bar{\mathbf{i}} \in \mathcal{I}$ since $\mathbf{p} \in \Lambda(\mathcal{I},\mathbf{m})$. Moreover, for any $\mathbf{i} \geq \bar{\mathbf{i}}$, $(\mathcal{L}_{k,i_k}-\mathcal{L}_{k,i_k-1})y_k^{p_k}=0$ by construction. Therefore,

$$\begin{aligned}
\mathcal{L}_{\mathcal{I}} \prod_{k=1}^{M} y_k^{p_k} &= \sum_{\mathbf{i}\in\mathcal{I}} \prod_{k=1}^{M} (\mathcal{L}_{k,i_k}-\mathcal{L}_{k,i_k-1}) y_k^{p_k} \\
&= \sum_{\mathbf{i}\leq\bar{\mathbf{i}}} \prod_{k=1}^{M} (\mathcal{L}_{k,i_k}-\mathcal{L}_{k,i_k-1}) y_k^{p_k} \\
&= \prod_{k=1}^{M} \sum_{i_k=1}^{\bar{i}_k} (\mathcal{L}_{k,i_k}-\mathcal{L}_{k,i_k-1}) y_k^{p_k} \\
&= \prod_{k=1}^{M} \mathcal{L}_{k,\bar{i}_k} y_k^{p_k}.
\end{aligned}$$

The third equality follows because the index set defined by

$$\mathbf{i} \in \mathbb{N}^M \quad \text{such that} \quad \mathbf{i} \leq \bar{\mathbf{i}}$$

corresponds to a (possibly anisotropic) tensor product rule. Finally, since

$$m_k(\bar{i}_k) = m_k(m_k^{-1}(p_k+1)) \geq p_k+1,$$

the one dimensional interpolant $\mathcal{L}_{k,\bar{i}_k}$ is exact for the monomial, $y_k^{p_k}$. This holds for all $k=1,\ldots,M$ and therefore, $\mathcal{L}_{\mathcal{I}}$ is exact on $\mathbb{P}(\Gamma,\mathcal{I},\mathbf{m})$. □

Barthelmann et al. [15] and Bäck et al. [14] prove similar results for the case when $\mathcal{I}$ represents classic isotropic and anisotropic Smolyak sparse grids.

4.2.3.2 Interpolation

The combination technique, (4.2.2), shows that the tensor product operator, $\mathcal{L}_{\mathcal{I}}$, has the specific form

$$(\mathcal{L}_{\mathcal{I}} f)(y) = \sum_{q=1}^{Q_{\mathcal{I}}} L_q(y) f(y_q)$$

where $Q_{\mathcal{I}} := |\mathcal{N}_{\mathcal{I}}|$, $\mathcal{N}_{\mathcal{I}} = \{y_q\}_{q=1}^{Q_{\mathcal{I}}}$ and $L_q \in \mathbb{P}(\Gamma, \mathcal{I}, \mathbf{m})$ for $q = 1, \ldots, Q_{\mathcal{I}}$. The goal of this section is to prove under certain circumstances that L_q is interpolatory, i.e. $L_q(y_s) = \delta_{qs}$ for $q, s = 1, \ldots, Q_{\mathcal{I}}$. In general, interpolation is not guaranteed for the operator, $\mathcal{L}_{\mathcal{I}}$. This fact can be seen by grouping nonzero coefficients, $c(\mathbf{i})$, in the combination technique form of $\mathcal{L}_{\mathcal{I}}$. It is clear to see that the operator $\mathcal{L}_{\mathcal{I}}$ is interpolatory for tensor product rules (i.e. $\mathcal{I} = \mathcal{I}_{\mathrm{TP}}(\ell, M)$) and Barthelmann et al. [15] prove this interpolation property for isotropic Smolyak rules (i.e. $\mathcal{I} = \mathcal{I}_{\mathrm{SM}}(\ell, M)$) as long as the one dimensional interpolation knots are nested. Similar results holds for anisotropic Smolyak [85]. These results can be generalized to the case of arbitrary admissible index sets, $\mathcal{I}$, as long as the one dimensional nodes are nested.

Proposition 4.2.5 *If $\mathcal{I} \subset \mathbb{N}^M$ is an admissible index set and*

$$\mathcal{N}_{k,j} \subset \mathcal{N}_{k,j+1} \quad \forall j > 0$$

for $k = 1, \ldots, M$, then $\mathcal{L}_{\mathcal{I}}$ is interpolatory; that is, for any $f \in C^0_\rho(\Gamma)$,

$$(\mathcal{L}_{\mathcal{I}} f)(y) = f(y) \quad \forall\, y \in \mathcal{N}_{\mathcal{I}}.$$

Proof: To prove this proposition, I will assume that $\mathcal{J} \subset \mathbb{N}^M$ is an admissible index set such that $\mathcal{L}_{\mathcal{J}}$ is interpolatory. I will then add an index, $\mathbf{j} \in \mathbb{N}^M$, to $\mathcal{J}$ such that $\mathcal{J} \cup \{\mathbf{j}\}$ remains admissible and prove that the resulting operator, $\mathcal{L}_{\mathcal{J}\cup\{\mathbf{j}\}}$, is also interpolatory.

Suppose $\mathcal{J} \subset \mathbb{N}^M$ is an admissible index set such that $\mathcal{L}_{\mathcal{J}}$ is interpolatory. Furthermore, suppose that $\mathcal{L}_{\mathcal{J}}$ is constructed using one dimensional interpolations operators built on nested interpolation knots

$$\mathcal{N}_{k,j} \subset \mathcal{N}_{k,j+1} \quad \forall j > 0 \quad \text{for } k = 1, \ldots, M.$$

Now, suppose the index $\mathbf{j} \in \mathbb{N}^M \setminus \mathcal{J}$ is such that $\mathcal{J} \cup \{\mathbf{j}\}$ is admissible. I will consider two cases: $\bar{y} \in \mathcal{N}_{\mathcal{J}}$ and $\bar{y} \in \mathcal{N}_{\mathbf{j}} \setminus \mathcal{N}_{\mathcal{J}}$, where $\mathcal{N}_{\mathbf{j}} := \mathcal{N}_{1,j_1} \times \cdots \times \mathcal{N}_{M,j_M}$ and $\mathcal{N}_{\mathcal{J}}$ is defined above. without loss of generality, assume

$$f = f_1 \cdot \ldots \cdot f_M \in C^0_{\rho_1}(\Gamma_1) \otimes \cdots \otimes C^0_{\rho_M}(\Gamma_M) \subset C^0_\rho(\Gamma).$$

For any $\bar{y} \in \mathcal{N}_{\mathcal{J}}$, it is easy to see from (4.2.1) that

$$(\mathcal{L}_{\mathcal{J}\cup\{\mathbf{j}\}} f)(\bar{y}) = (\mathcal{L}_{\mathcal{J}} f)(\bar{y}) + (\Delta_{\mathbf{j}} f)(\bar{y}) = f(\bar{y}) + \prod_{k=1}^{M} (\Delta_{k,j_k} f_k)(\bar{y}_k)$$

since $\mathcal{L}_{\mathcal{J}}$ is interpolatory. Since $\mathbf{j} \notin \mathcal{J}$ and $\mathcal{J} \cup \{\mathbf{j}\}$ is admissible, there exists a $k \in \{1, \ldots, M\}$ such that $j_k > i_k$ for all $\mathbf{i} \in \mathcal{J}$. By definition of $\mathcal{N}_{\mathcal{J}}$, $\bar{y}_k \in \mathcal{N}_{k,i}$ for at least one $i < j_k$. Therefore, the nestedness of the one dimensional nodes implies $\bar{y}_k \in \mathcal{N}_{k,j_k-1} \subset \mathcal{N}_{k,j_k}$. This shows that $\Delta_{j_k} f_k(\bar{y}_k) = 0$ and hence

$$(\mathcal{L}_{\mathcal{J}\cup\{\mathbf{j}\}} f)(\bar{y}) = f(\bar{y}) + \prod_{k=1}^{M} (\Delta_{k,j_k} f_k)(\bar{y}_k) = f(\bar{y}),$$

proving that $\mathcal{L}_{\mathcal{J}\cup\{\mathbf{j}\}}$ is interpolatory on the sparse grid, $\mathcal{N}_{\mathcal{J}}$.

On the other hand, for any $\bar{y} \in \mathcal{N}_{\mathbf{j}} \setminus \mathcal{N}_{\mathcal{J}}$. Define the set

$$\mathcal{J}_{\mathbf{j}} := \left\{ \mathbf{i} \in \mathbb{N}^M \ : \ \mathbf{i} \leq \mathbf{j} \right\}.$$

By admissibility, $\mathcal{J}_{\mathbf{j}} \subset \mathcal{J} \cup \{\mathbf{j}\}$ and $\mathcal{J}_{\mathbf{j}} \setminus \{\mathbf{j}\} \subset \mathcal{J}$. With this definition, one can write

$$(\mathcal{L}_{\mathcal{J}\cup\{\mathbf{j}\}} f)(\bar{y}) = (\mathcal{L}_{\mathcal{J}_{\mathbf{j}}} f)(\bar{y}) + (\mathcal{L}_{\mathcal{J}\setminus\mathcal{J}_{\mathbf{j}}} f)(\bar{y}).$$

Notice that $\mathcal{L}_{\mathcal{J}_{\mathbf{j}}}$ defines the anisotropic tensor product operator, $\mathcal{L}_{\mathbf{j}}$, which is interpolatory. Furthermore, notice that

$$\mathcal{J} \setminus \mathcal{J}_{\mathbf{j}} = \{\mathbf{i} \in \mathcal{J} \ : \ \exists\, k \in \{1, \ldots, M\} \text{ such that } i_k > j_k\}.$$

Therefore, given an index $\mathbf{i} \in \mathcal{J} \setminus \mathcal{J}_{\mathbf{j}}$, there exists $k \in \{1, \ldots, M\}$ such that $i_k > j_k$. By nestedness of the one dimensional nodes, it follows that $\bar{y}_k \in \mathcal{N}_{k,j_k} \subset \mathcal{N}_{k,i_k-1} \subset \mathcal{N}_{k,i_k}$. This implies that $(\Delta_{\mathbf{i}} f)(\bar{y}) = 0$ for all $\mathbf{i} \in \mathcal{J} \setminus \mathcal{J}_{\mathbf{j}}$. Hence,

$$(\mathcal{L}_{\mathcal{J}\cup\{\mathbf{j}\}} f)(\bar{y}) = (\mathcal{L}_{\mathcal{J}_{\mathbf{j}}} f)(\bar{y}) + (\mathcal{L}_{\mathcal{J}\setminus\mathcal{J}_{\mathbf{j}}} f)(\bar{y}) = (\mathcal{L}_{\mathbf{j}} f)(\bar{y}) = f(\bar{y}),$$

and thus completing the proof of Proposition 4.2.5. □

4.2.3.3 Approximation Error

The previous results described the polynomial exactness and interpolation properties of the operator, $\mathcal{L}_{\mathcal{I}}$, for admissible index sets, $\mathcal{I}$. My goal, now, is to give an explicit representation of the interpolation error associated with $\mathcal{L}_{\mathcal{I}}$. First, I will present the necessary notion for the results to follow. Secondly, I will prove a lemma which gives a specific form for the interpolation error. Finally, I will extend this specific error form to the case of interpolating functions with analytic extensions one tensor products of one dimensional Clenshaw-Curtis knots.

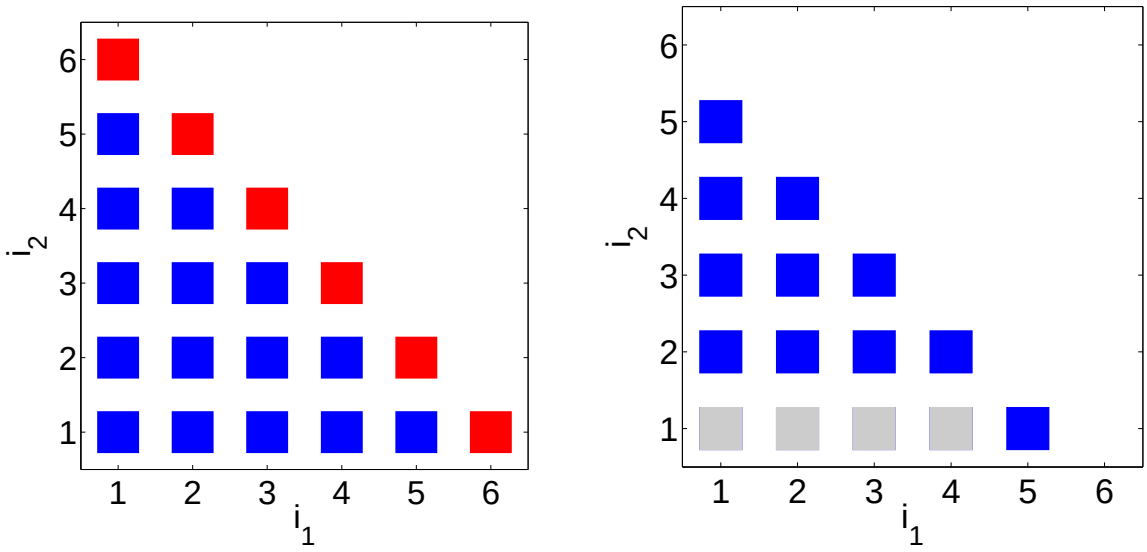


Figure 4.4: The left image contains the level 4 isotropic Smolyak index set, $\mathcal{I}$, (blue) and corresponding forward margin, $\mathcal{M}(\mathcal{I})$ (red). The right image depicts $\mathcal{I}_M = \mathcal{I}$ with $M = 2$ (blue and grey) and the recursively defined $\mathcal{I}_1$ (grey).

The results in this section are highly dependent on the notion of the forward margin of the admissible index set, $\mathcal{I}$. I will denote this forward margin by

$$\mathcal{M}(\mathcal{I}) := \{\mathbf{i} + \mathbf{e}_k \notin \mathcal{I} \ : \ k = 1, \ldots, M\}$$

where $\mathbf{e}_k$ denotes the index with 1 in the k^{th} position and zeros everywhere else. See Figure 4.4 for an example of the forward margin. Additionally, the results will be

dependent on the following recursively defined sequence of sets

$$\mathcal{I}_M := \mathcal{I} \quad \text{and} \quad \mathcal{I}_d := \left\{ \mathbf{i} \in \mathbb{N}^d \ : \ \exists \, i_{d+1} \in \mathbb{N} \text{ such that } (\mathbf{i}, i_{d+1}) \in \mathcal{I}_{d+1} \right\}$$

for $d = M - 1 \ldots, 1$ (e.g. see Figure 4.4 for a two dimensional example of these recursively defined index sets). Since $\mathcal{I}$ is an admissible subset of $\mathbb{N}^M$, $\mathcal{I}_d$ is an admissible subset of $\mathbb{N}^d$. Furthermore, I will use the functions $\pi_d : \mathcal{I}_d \to \mathbb{N}$ defined by

$$\pi_d(\mathbf{i}) := \max \left\{ i_{d+1} \in \mathbb{N} \ : \ (\mathbf{i}, i_{d+1}) \in \mathcal{I}_{d+1} \right\}.$$

From this definition, it is clear that for all $\mathbf{i} \in \mathcal{I}_d$, the index, $\mathbf{j} = (\mathbf{i}, i_d) \in \mathcal{I}_{d+1}$ for all $i_d \leq \pi_d(\mathbf{i})$. The following result is a generalization of arguments in the proof of Lemma 3.2 in [85]. The result in [85], is specifically for anisotropic Smolyak sparse grid operators built on the index sets

$$\mathcal{I} = X_\alpha(\ell, M) = \left\{ \mathbf{i} \in \mathbb{N}^M \ : \ \sum_{k=1}^{M} (i_k - 1)\alpha_k \leq \ell \bar{\alpha} \right\}$$

where $\bar{\alpha} := \min_{1 \leq k \leq M} \alpha_k$. The corresponding sets, $\mathcal{I}_d$, are denoted by $\mathcal{I}_d = X_\alpha(\ell, d)$, the margins are denoted $\mathcal{M}(\mathcal{I}_d) = \tilde{X}_\alpha(\ell, d)$, and the mapping π_d has the particular form,

$$\pi_d(\mathbf{i}) = \left\lfloor 1 + \ell \frac{\bar{\alpha}}{\alpha_{d+1}} - \sum_{k=1}^{d} (i_k - 1) \frac{\alpha_k}{\alpha_{d+1}} \right\rfloor$$

where $\lfloor \cdot \rfloor$ denotes the floor of a real number. The following notation will be used in the statement and proof of Lemma 4.2.6. Let $\mathbf{I}_d$ denote the identity operator on $L^2_{\rho_d}(\Gamma_d)$ and $\widehat{\mathbf{I}}_d$ denote the identity operator acting on $L^2_{\widehat{\rho}_d}(\widehat{\Gamma}_d)$ where $\widehat{\Gamma}_d := \prod_{k=1}^{d} \Gamma_k$ and $\widehat{\rho}_d = \prod_{k=1}^{d} \rho_k$ for $d = 1, \ldots, M$, i.e. $\widehat{\mathbf{I}}_d = \bigotimes_{k=1}^{d} \mathbf{I}_k$.

Lemma 4.2.6 *The error associated with the tensor product interpolation operator, $\mathcal{L}_\mathcal{I}$, built on an admissible index set, $\mathcal{I} \subset \mathbb{N}^M$, has the particular form*

$$(\widehat{\mathbf{I}}_M - \mathcal{L}_\mathcal{I}) = \sum_{k=1}^{M} (\mathbf{R}_k \otimes \widehat{\mathbf{I}}_{M-k})$$

where the operators, $\mathbf{R}_k$*, are defined as*

$$\mathbf{R}_k := \sum_{\mathbf{i}\in\mathcal{I}_{d-1}} \left\{ \bigotimes_{k=1}^{d-1} \Delta_{k,i_k} \right\} \otimes (\mathbf{I}_d - \mathcal{L}_{d,\widehat{i}_d-1})$$

and $\widehat{i}_d = \pi_{d-1}(\mathbf{i}) + 1$*.*

Proof: By admissibility of $\mathcal{I}$, the index $(\mathbf{i}, i_M) \in \mathcal{I}$ for all $1 \leq k \leq \pi_{M-1}(\mathbf{i})$; therefore, the following decomposition of $\mathcal{L}_{\mathcal{I}}$ is valid

$$\begin{aligned}
\mathcal{L}_{\mathcal{I}} &= \sum_{\mathbf{i}\in\mathcal{I}} \bigotimes_{k=1}^{M} \Delta_{k,i_k} \\
&= \sum_{\mathbf{i}\in\mathcal{I}_{M-1}} \left\{ \bigotimes_{k=1}^{M-1} \Delta_{k,i_k} \right\} \otimes \left\{ \sum_{d=1}^{\pi_{M-1}(\mathbf{i})} \Delta_{M,d} \right\} \\
&= \sum_{\mathbf{i}\in\mathcal{I}_{M-1}} \left\{ \bigotimes_{k=1}^{M-1} \Delta_{k,i_k} \right\} \otimes \mathcal{L}_{M,\pi_{M-1}(\mathbf{i})} \qquad (4.2.5)
\end{aligned}$$

Now, plugging the decomposition (4.2.5) into the tensor product interpolation error gives

$$\begin{aligned}
\widehat{\mathbf{I}}_M - \mathcal{L}_{\mathcal{I}} &= \widehat{\mathbf{I}}_M - \sum_{\mathbf{i}\in\mathcal{I}_{M-1}} \left\{ \bigotimes_{k=1}^{M-1} \Delta_{k,i_k} \right\} \otimes \left\{ \mathcal{L}_{M,\pi_{M-1}(\mathbf{i})} - \mathbf{I}_M \right\} \\
&\qquad - \sum_{\mathbf{i}\in\mathcal{I}_{M-1}} \left\{ \bigotimes_{k=1}^{M-1} \Delta_{k,i_k} \right\} \otimes \mathbf{I}_M \\
&= \sum_{\mathbf{i}\in\mathcal{I}_{M-1}} \left\{ \bigotimes_{k=1}^{M-1} \Delta_{k,i_k} \right\} \otimes \left\{ \mathbf{I}_M - \mathcal{L}_{M,\pi_{M-1}(\mathbf{i})} \right\} + (\widehat{\mathbf{I}}_{M-1} - \mathcal{L}_{\mathcal{I}_{M-1}}) \otimes \mathbf{I}_M \\
&= \sum_{k=2}^{M} (\mathbf{R}_k \otimes \widehat{\mathbf{I}}_{M-k}) + (\mathbf{I}_1 - \mathcal{L}_{\mathcal{I}_1}) \otimes \widehat{\mathbf{I}}_{M-1}.
\end{aligned}$$

Here, $\mathbf{R}_d$ is defined as

$$\mathbf{R}_d := \sum_{\mathbf{i}\in\mathcal{I}_{d-1}} \left\{ \bigotimes_{k=1}^{d-1} \Delta_{k,i_k} \right\} \otimes (\mathbf{I}_d - \mathcal{L}_{d,\widehat{i}_d-1})$$

and $\widehat{i}_d = \pi_{d-1}(\mathbf{i}) + 1$. This proves the desired result since $\mathbf{R}_1 = (\mathbf{I}_1 - \mathcal{L}_{\mathcal{I}_1})$. □

I will now use Lemma 4.2.6 to prove an upper bound on the error of interpolating functions $f \in C^0_\rho(\Gamma)$ which admit an analytic extension in each direction, Γ_k, on the elliptic discs, $D_{r_k}(\Gamma_k)$, using the tensor product operator, $\mathcal{L}_\mathcal{I}$, built on one dimensional Clenshaw-Curtis interpolation knots. Of course, these results are valid only for $\Gamma_k = [a_k, b_k]$ with $-\infty < a_k < b_k < \infty$, for $k = 1, \ldots, M$, endowed with a uniform distribution. This result is a generalization of Lemma 3.2 in [85].

Corollary 4.2.7 *Suppose $f \in L^2_\rho(\Gamma)$ has an analytic extension in each direction, Γ_k, on the elliptic discs, $D_{r_k}(\Gamma_k)$. Furthermore, suppose $\mathcal{L}_\mathcal{I}$ is built on one dimensional Clenshaw-Curtis interpolation knots. Then the error associated with the tensor product interpolation operator, $\mathcal{L}_\mathcal{I}$, satisfies the upper bound*

$$\|f - \mathcal{L}_\mathcal{I} f\|_{L^\infty_\rho(\Gamma)} \leq \sum_{k=1}^{M} R_k$$

where

$$R_k := \sum_{\mathbf{i} \in \mathcal{M}(\mathcal{I}_k)} C D^{k-1} \left(\prod_{d=1}^{k} i_d \right) e^{-h(\mathbf{i},k)} \quad \textit{and} \quad h(\mathbf{i},k) = \sum_{d=1}^{k} \log(r_d) 2^{i_d - 1},$$

and C, D are positive constants defined in (4.1.4) *and* (4.1.5).

Proof: By Lemma 4.2.6, the interpolation error satisfies

$$(\widehat{\mathbf{I}}_M - \mathcal{L}_\mathcal{I}) = \sum_{k=1}^{M} (\mathbf{R}_k \otimes \widehat{I}_{M-k}).$$

The principal task is to bound the norms of $\mathbf{R}_d$. First, notice that $(\mathbf{i}, \widehat{i}_d) \in \mathcal{M}(\mathcal{I}_d)$ for all $\mathbf{i} \in \mathcal{I}_{d-1}$. Now,

$$\begin{aligned}
\|\mathbf{R}_d f\|_{C^0_\rho(\Gamma)} &\leq \sum_{\mathbf{i} \in \mathcal{I}_{d-1}} \left\{ \prod_{k=1}^{d-1} \|\Delta_{k,i_k} f\|_{C^0_\rho(\Gamma)} \right\} \cdot \|f - \mathcal{L}_{d,\widehat{i}_d - 1} f\|_{C^0_\rho(\Gamma)} \\
&\leq \sum_{\mathbf{i} \in \mathcal{I}_{d-1}} C D^{d-1} \left(\prod_{k=1}^{d-1} i_k \right) (\widehat{i}_d - 1) \exp\Big(- \sum_{k=1}^{d-1} \log(r_k) 2^{i_k - 1} - \log(r_d) 2^{\widehat{i}_d - 1} \Big) \\
&\leq \sum_{\mathbf{i} \in \mathcal{M}(\mathcal{I}_d)} C D^{d-1} \left(\prod_{k=1}^{d} i_k \right) e^{-h(\mathbf{i},d)} =: R_d.
\end{aligned}$$

This proves the desired result since $\mathbf{R}_1 = (\mathbf{I}_1 - \mathcal{L}_{\mathcal{I}_1})$ and $R_1 := \sum_{i \in \mathcal{M}(\mathcal{I}_1)} C i e^{-\log(r_1) 2^{i-1}}$. $\square$

Corollary 4.2.7 shows that the error associated with the tensor product interpolation operator, $\mathcal{L}_{\mathcal{I}}$, is concentrated on the margin of $\mathcal{I}$. Therefore, in the adaptive selection of an index set, $\mathcal{I}$, it is sufficient to control the error on the margin of that index set, $\mathcal{M}(\mathcal{I})$. In this case, the convergence of Algorithm 4.2.3 holds when using the error indicators

$$\eta_{\mathbf{i}} = C D^{d-1} \left(\prod_{k=1}^{d} i_k \right) e^{-h(\mathbf{i},d)}$$

for all indices $\mathbf{i} \in \mathcal{M}(\mathcal{I}_d)$. In general, this error bound is not computable since one often does not know the semi-axis lengths of the ellipses in $\mathbb{C}$ for which a function has an analytic extension. On the other hand, one can use the results in Corollary 4.2.7 to prove usable error bounds for specific choices of index set, $\mathcal{I}$. In Corollary 4.2.8, I restate Theorem 3.4 from [85]. This result uses Corollary 4.2.7 to prove an error bound for the anisotropic Smolyak sparse grid based on the index set

$$\mathcal{I} = X_\alpha(\ell, M) = \left\{ \mathbf{i} \in \mathbb{N}^M \ : \ \sum_{k=1}^{M} (i_k - 1)\alpha_k \leq \ell \bar{\alpha} \right\}.$$

To clearly state this result, I will employ the following notation:

$$\bar{\alpha} := \min_{k=1,\ldots,M} \alpha_k \quad \text{and} \quad \mathcal{A} := \sum_{k=1}^{M} \alpha_k.$$

Corollary 4.2.8 (Theorem 3.4 in [85]) *Suppose $f \in L^2_\rho(\Gamma)$ has an analytic extension in each direction, Γ_k, on the elliptic discs, $D_{r_k}(\Gamma_k)$. Furthermore, suppose $\mathcal{L}_{\mathcal{I}}$ is built on one dimensional Clenshaw-Curtis interpolation knots and $\mathcal{I}$ corresponds to the anisotropic Smolyak index set*

$$\mathcal{I} = X_\alpha(\ell, M) = \left\{ \mathbf{i} \in \mathbb{N}^M \ : \ \sum_{k=1}^{M} (i_k - 1)\alpha_k \leq \ell \bar{\alpha} \right\}$$

with $\alpha_k = \log\left(\frac{2r_k}{|\Gamma_k|} + \sqrt{1 + \frac{4r_k^2}{|\Gamma_k|^2}} \right)$. Then, the interpolation error associated with $\mathcal{L}_{\mathcal{I}}$ is bounded by

$$\| f - \mathcal{L}_{\mathcal{I}} f \|_{L^\infty_\rho(\Gamma)} \leq C(\alpha, M)\, e^{\ell - \mu(\ell, M)}$$

where

$$\mu(\ell, N) := \begin{cases} \frac{\ell\, e\, \log(2)\, \bar{\alpha}}{2} & \text{if } 0 \leq \ell \leq \frac{\mathcal{A}}{\bar{\alpha}\log(2)} \\ \frac{\mathcal{A}}{2} 2^{\frac{\ell\bar{\alpha}}{\mathcal{A}}} & \text{otherwise} \end{cases}$$

and the constant $C = C(\alpha, M)$ is independent of the level ℓ.

Proof: This result follows from Corollary 4.2.7 and Lemma 3.3 in [85]. □

The authors of [85] extend the error bound in Corollary 4.2.8 to be dependent on the number of interpolation knots. In Theorem 3.8 of [85], the authors prove that under the conditions of Corollary 4.2.8, the interpolation error associated with $\mathcal{L}_{\mathcal{I}}$ satisfy

$$\|f - \mathcal{L}_{\mathcal{I}} f\|_{L^{\infty}_{\rho}(\Gamma)} \leq C(\alpha, M)\, Q^{-\nu} \tag{4.2.6}$$

where $\nu := \frac{\bar{\alpha}\log(2)e-1}{log(2)+\sum_{k=1}^{M}\frac{\bar{\alpha}}{\alpha_k}}$ whenever $0 \leq \ell \leq \frac{\mathcal{A}}{\bar{\alpha}\log(2)}$. The reader will recall that the error bound (4.2.6) is exactly the bound required by Assumption 3.2.4. Furthermore, in the case that $\frac{\mathcal{A}}{\bar{\alpha}\log(2)} < \ell$, one gets sub-exponential convergence of the sparse grid algorithm (c.f. equation 3.23 in [85]).

Remark 4.2.9 *Similar error bounds hold for anisotropic Smolyak rules built on Gaussian interpolation knots (see Theorem 3.13 in [85]). Moreover, similar error bounds hold for isotropic Smolyak and tensor product rules built on Clenshaw-Curtis and Gaussian interpolation knots (see Theorem 6.2 in [9] for convergence of isotropic Smolyak and Theorem 4.1 in [9] for convergence of tensor product rules).*

Chapter 5

Trust Regions and Adaptivity

In this chapter, I develop a framework for the adaptive solution of optimization governed by PDEs with uncertain coefficients. This framework is built on the retrospective trust region algorithm. I prove that, with inexact gradient information, this modified trust region algorithm remains globally convergent. In the trust region framework, I use inexact gradient bounds and *a posteriori* error indicators to guide model adaptation. For optimization of PDEs with uncertain coefficients, one requires error indicators for the finite element method, the sparse grids used in the stochastic collocation method, and the model reduction basis in the case of time dependent problems.

The trust region method is a very popular and powerful optimization framework for solving general nonlinear programming problems [81, 82, 94, 109, 41]. The trust region framework offers quite a bit of flexibility in exactness of function evaluations and gradients. In fact, one can prove global convergence of the trust region method when the function evaluations are inexact [41], as well as, when the gradient evaluations are inexact [36]. This flexibility makes the trust region framework an ideal candidate for a general model adaptation framework for solving PDE constrained optimization problems. Much work has gone into trust region frameworks for managing approximate models in optimization [43, 44, 3]. This model management framework

is the basis of my adaptive framework.

5.1 The Basic Trust Region Algorithm

In this section, I will formulate the basic trust region algorithm for unconstrained optimization in a Hilbert space. Let $\mathcal{Z}$ be a Hilbert space. The basic trust region algorithm for solving the nonlinear programming problem

$$\min_{z\in\mathcal{Z}} \widehat{J}(z)$$

seeks to compute steps, $z_{k+1} = z_k + s_k \in \mathcal{Z}$, by solving an "inexpensive" subproblem. In the classic trust region theory, inexpensive subproblems are constructed using second order Taylor approximations of $\widehat{J}(z)$ centered around the current iterate, z_k. Taylor approximations are, in general, accurate in a small ball containing the iterate, z_k. Such a ball gives rise to the notion of a trust region and the basic trust region subproblem is

$$\min_{s\in\mathcal{Z}} m_k(s) \quad \text{subject to} \quad \|s\|_{\mathcal{Z}} \leq \Delta_k \tag{5.1.1}$$

for some inexpensive model, $m_k : \mathcal{Z} \to \mathbb{R}$ and $m_k(s) \approx \widehat{J}(z_k + s)$, and some positive scalar, $\Delta_k > 0$. The steps computed by solving the subproblem need not be exact minimizers. In general, (5.1.1) need not have a solution since the set

$$\mathcal{B}_k := \{s \in \mathcal{Z} \ : \ \|s\|_{\mathcal{Z}} \leq \Delta_k\}$$

may not be compact for a general Hilbert space, $\mathcal{Z}$ (c.f. page 275 in [41]). The trust region theory only requires the solution of the subproblem to satisfy the *fraction of Cauchy decrease* condition

$$m_k(0) - m_k(s_k) \geq \kappa_0 \|\nabla m_k(0)\|_{\mathcal{Z}} \min\left\{\Delta_k, \ \frac{\|\nabla m_k(0)\|_{\mathcal{Z}}}{\beta_k}\right\} \tag{5.1.2}$$

where $\kappa_0 \in (0,1)$ and $\beta_k = 1 + \sup_{s\in\mathcal{B}_k} \|\nabla^2 m_k(s)\|_{\mathcal{L}(\mathcal{Z},\mathcal{Z})}$. The fraction of Cauchy decrease condition (5.1.2) ensures that the decrease in the modeled objective function,

$m_k(s)$, corresponding to the approximate solution of (5.1.1) is at least as large as the decreases of $m_k(s)$ corresponding to the minimizer of (5.1.1) in the negative gradient $(-\nabla m_k(0))$ direction, i.e. the *Cauchy point*. The basic trust region algorithm is stated in Algorithm 5.1.1.

Algorithm 5.1.1 - Basic Trust Region Algorithm:

1. ***Initialization:*** *Given* m_k, z_k, Δ_k, $\gamma_1 < 1 < \gamma_2$, *and* $0 < \eta_1 < \eta_2 < 1$.

2. ***Step Computation:*** *Approximate a solution,* s_k, *to the trust region subproblem* (5.1.1) *which satisfies condition* (5.1.2).

3. ***Step Acceptance:*** *Compute the ratio* $\rho_k = \frac{\text{ared}_k}{\text{pred}_k}$ *where*

$$\text{pred}_k = m_k(0) - m_k(s_k) \quad \textit{and} \quad \text{ared}_k = \widehat{J}(z_k) - \widehat{J}(z_k + s_k).$$

if $\rho_k \geq \eta_1$ ***then*** $z_{k+1} = z_k + s_k$ ***else*** $z_{k+1} = z_k$ ***end if***

4. ***Trust Region Radius Update:***

if $\rho_k \leq \eta_1$ ***then*** $\Delta_{k+1} \in (0, \gamma_1 \|s_k\|_{\mathcal{Z}}]$ ***end if***

if $\rho_k \in (\eta_1, \eta_2)$ ***then*** $\Delta_{k+1} \in [\gamma_1 \|s_k\|_{\mathcal{Z}}, \Delta_k]$ ***end if***

if $\rho_k \geq \eta_2$ ***then*** $\Delta_{k+1} \in [\Delta_k, \gamma_2 \Delta_k]$ ***end if***

5. ***Model Update:*** *Choose a new model* $m_{k+1}(s)$.

The standard assumptions on the objective function, $\widehat{J}(z)$, and the successive approximations of the objective function, $m_k(s)$, which guarantee first order convergence of Algorithm 5.1.1 are

Assumptions 5.1.2 - Basic Trust Region:

- $\widehat{J}$ *is twice continuously Fréchet differentiable and bounded below;*

- *m_k is twice continuously Fréchet differentiable for $k = 1, 2, \ldots$;*
- *There exists $\kappa_1 > 0$ such that $\|\nabla^2 \widehat{J}(z)\|_{\mathcal{L}(\mathcal{Z},\mathcal{Z})} \leq \kappa_1$ for all $z \in \mathcal{Z}$;*
- *There exists $\kappa_2 \geq 1$ such that $\|\nabla^2 m_k(s)\|_{\mathcal{L}(\mathcal{Z},\mathcal{Z})} \leq \kappa_2 - 1$ for all $z \in \mathcal{Z}$ and for all $k = 1, 2, \ldots$;*
- *There exists $\xi > 0$ independent of k such that*

$$\|\nabla m_k(0) - \nabla \widehat{J}(z_k)\|_{\mathcal{Z}} \leq \xi \min\{\|\nabla m_k(0)\|_{\mathcal{Z}}, \Delta_k\} \tag{5.1.3}$$

for $k = 1, 2, \ldots$.

The inexact gradient condition, (5.1.3), is due to Heinkenschloss and Vicente [60]. This condition is slightly stronger than the classic condition due to Carter [36]

$$\|\nabla m_k(0) - \nabla \widehat{J}(z_k)\|_{\mathcal{Z}} \leq \xi \|\nabla m_k(0)\|_{\mathcal{Z}}$$

with $\xi < 1 - \eta_2$. The downside of this gradient condition is that ξ is required to be less than 1. In many practical applications, one does not know exactly the scaling coefficients in front of their error bounds. Thus, (5.1.3) is preferable to compute with since one does not need to know ξ exactly. As stated, these assumptions are sufficient to show that Algorithm 5.1.1 converges to a first order stationary point (c.f. see [81, 41]).

5.2 The Retrospective Trust Region

In Algorithm 5.1.1, the trust region radius, Δ_k, is always updated according to the current model, $m_k(s)$; hence, the new trust region radius, Δ_{k+1}, may be insufficient to handle the new model, $m_{k+1}(s)$. Bastin, et al. created the retrospective trust region algorithm to circumvent this possible pitfall [17]. In the retrospective framework, steps are accepted according to the performance index

$$\rho_k = \frac{\text{ared}_k}{\text{pred}_k} = \frac{\widehat{J}(z_k) - \widehat{J}(z_k + s_k)}{m_k(0) - m_k(s_k)}$$

as in Algorithm 5.1.1, but the trust region radius is updated retrospectively according to the new model. Once a step is accepted according to ρ_k, the model is updated to $m_{k+1}(s)$ and one computes the retrospective performance index

$$\tilde{\rho}_k = \frac{\widehat{J}(z_k) - \widehat{J}(z_k + s_k)}{m_{k+1}(0) - m_{k+1}(s_k)}.$$

Using this index, $\tilde{\rho}_k$, the trust region radius is fit to the new model, $m_{k+1}(s)$. As with Algorithm 5.1.1, one need not solve the trust region subproblem exactly. The approximate solution to the trust region subproblem must satisfy the fraction of Cauchy decrease condition (5.1.2). The retrospective trust region algorithm is stated in Algorithm 5.2.1.

Algorithm 5.2.1 - Retrospective Trust Region Algorithm:

1. ***Initialization:*** *Given* m_k, z_k, Δ_k, $0 < \gamma_1 \leq \gamma_2 < 1$, $\Delta_{\max} > 0$, $0 < \eta_0 < 1$, *and* $0 < \eta_1 < \eta_2 < 1$.

2. ***Step Computation:*** *Approximate a solution,* s_k, *to the trust region subproblem* (5.1.1) *which satisfies* (5.1.2).

3. ***Step Acceptance:*** *Compute the ratio* $\rho_k = \frac{\text{ared}_k}{\text{pred}_k}$ *where*

$$\text{pred}_k = m_k(0) - m_k(s_k) \quad \textit{and} \quad \text{ared}_k = \widehat{J}(z_k) - \widehat{J}(z_k + s_k).$$

   if $\rho_k \geq \eta_0$ ***then*** $z_{k+1} = z_k + s_k$ ***else*** $z_{k+1} = z_k$ ***end if***

4. ***Model Update:*** *Choose a new model,* m_{k+1}.

5. ***Trust Region Radius Update:***

   if $z_{k+1} = z_k$ ***then*** $\Delta_{k+1} \in (0, \gamma_1 \|s_k\|_{\mathcal{Z}}]$

   else *define*

$$\tilde{\rho}_{k+1} = \frac{\widehat{J}(z_{k+1}) - \widehat{J}(z_k)}{m_{k+1}(0) - m_{k+1}(-s_k)}$$

and update Δ_{k+1} *by*

if $\tilde{\rho}_{k+1} \leq \eta_1$	***then***	$\Delta_{k+1} \in (0, \gamma_2 \|s_k\|_{\mathcal{Z}}]$	***end if***
if $\tilde{\rho}_{k+1} \in (\eta_1, \eta_2)$	***then***	$\Delta_{k+1} \in [\gamma_2 \|s_k\|_{\mathcal{Z}}, \Delta_k]$	***end if***
if $\tilde{\rho}_{k+1} \geq \eta_2$	***then***	$\Delta_{k+1} \in [\Delta_k, \Delta_{\max}]$	***end if***

The assumptions I will make on the objective function, $\widehat{J}(z)$, and the successive approximations to the objective function, $m_k(s)$, to guarantee first order convergence are

Assumptions 5.2.2 - Retrospective Trust Region:

- $\widehat{J}$ *is twice continuously Fréchet differentiable and bounded below;*
- m_k *is twice continuously Fréchet differentiable for* $k = 1, 2, \ldots$*;*
- *There exists* $\kappa_1 > 0$ *such that* $\|\nabla^2 \widehat{J}(z)\|_{\mathcal{L}(\mathcal{Z},\mathcal{Z})} \leq \kappa_1$ *for all* $z \in \mathcal{Z}$*;*
- *There exists* $\kappa_2 \geq 1$ *such that* $\|\nabla^2 m_k(s)\|_{\mathcal{L}(\mathcal{Z},\mathcal{Z})} \leq \kappa_2 - 1$ *for all* $s \in \mathcal{Z}$ *and for all* $k = 1, 2, \ldots$*;*
- *There exists* $\xi > 0$ *independent of* k *such that*

$$\|\nabla m_k(0) - \nabla \widehat{J}(z_k)\|_{\mathcal{Z}} \leq \xi \min\{\|\nabla m_k(0)\|_{\mathcal{Z}}, \Delta_{k-1}\} \tag{5.2.1}$$

for $k = 1, 2, \ldots$.

These assumptions are relaxed from the assumptions made in the original retrospective trust region work [17]. Namely, Bastin, et al. assume that the model at $s = 0$, $m_k(0)$, and its gradient at $s = 0$, $\nabla m_k(0)$, are equal to the objective function at z_k, $\widehat{J}(z_k)$, and its gradient at z_k, $\nabla \widehat{J}(z_k)$, respectively. I will prove later that my weaker assumptions are sufficient to prove that the retrospective trust region algorithm converges to a first order critical point.

Remark 5.2.3 *As seen in Algorithm 5.2.2, each successful step requires two new model evaluations to compute $\tilde{\rho}_{k+1}$ (i.e. $m_{k+1}(0)$ and $m_{k+1}(-s_k)$). For the optimization problems considered in this thesis, additional model evaluations are very expensive as they require the stochastic collocation solution to the state equation. Hence, for some classes of optimization problems it is unclear whether or not the additional computational cost of the retrospective trust region algorithm is worth while.*

5.2.1 Discussion of Stopping Criterion

As a stopping criterion, I suggest using a model gradient stopping test and a step size stopping test. First of all, if the computed step, s_k is "sufficiently" small (sufficient here depends on the scaling of the problem), then the trust region algorithm is not making significant progress and should be terminated. Given a step tolerance stol > 0, this condition reads

$$\|s_k\|_{\mathcal{Z}} \leq \text{stol}.$$

On the other hand, if the modeled gradient, $\nabla m_k(0)$, is "sufficiently" small (again sufficient depends on the scaling of the problem), then the algorithm should be terminated. Moreover, the inexact gradient conditions (5.1.3) implies that

$$\|\nabla \widehat{J}(z_k)\|_{\mathcal{Z}} \leq \|\nabla \widehat{J}(z_k) - \nabla m_k(0)\|_{\mathcal{Z}} + \|\nabla m_k(0)\|_{\mathcal{Z}} \leq (1+\xi)\|\nabla m_k(0)\|_{\mathcal{Z}}.$$

Thus, if gtol > 0 and $\|\nabla m_k(0)\|_{\mathcal{Z}} <$ gtol, then it is clear that

$$\|\nabla \widehat{J}(z_k)\|_{\mathcal{Z}} \leq (1+\xi)\text{gtol}.$$

5.2.2 Convergence of the Retrospective Trust Region

In this section, I prove that under Assumptions 5.2.2, the Algorithm 5.2.1 converges to a first order critical point. First, I prove that the sequence of trust region radii must converge to zero if the norm of the gradients are bounded away from zero. Secondly, I show that under these assumptions, ρ_k and $\tilde{\rho}_{k+1}$ converge to one. Finally,

these results are combined to prove that Algorithm 5.2.2 converges to a first order critical point. Most results presented here follow the standard convergence proof for the basic trust region algorithm provided in Theorem 4.10 in [81], although care must be taken to handle the retrospective trust region update.

Lemma 5.2.4 *Suppose there exists $\epsilon > 0$ such that $\|\nabla m_k(0)\|_{\mathcal{Z}} \geq \epsilon$ for k sufficiently large. Then the sequence of trust region radii, $\{\Delta_k\}$, produced by Algorithm 5.2.1 satisfies*

$$\sum_{k=1}^{\infty} \Delta_k < \infty.$$

Proof: First notice that the result of the theorem holds if there is only a finite number of successful iterations because for sufficiently large k, $\Delta_{k+1} \leq \gamma_1 \Delta_k$. Now, if there is an infinite sequence of successful iterations $\{k_i\}$ then for sufficiently large i the fraction of Cauchy decrease condition (5.1.2) implies

$$\begin{aligned}
\widehat{J}(z_{k_i}) - \widehat{J}(z_{k_{i+1}}) \geq & \widehat{J}(z_k) - \widehat{J}(z_{k+1}) \\
\geq & \eta_0 (m_k(0) - m_k(s_k)) \\
\geq & \eta_0 \kappa_0 \|\nabla m_k(0)\|_{\mathcal{Z}} \min \left\{ \Delta_k, \frac{\|\nabla m_k(0)\|_{\mathcal{Z}}}{\beta_k} \right\} \\
\geq & \eta_0 \kappa_0 \Delta_k \epsilon.
\end{aligned}$$

This implies that $\sum_{i=1}^{\infty} \Delta_{k_i} < \infty$. Furthermore, for every unsuccessful iteration $k \notin \{k_i\}$, the trust region radius satisfies $\Delta_k \leq \gamma_1^{k-k_j} \Delta_{k_j}$ where $k_j \in \{k_i\}$ is the largest index such that $k_j < k$. The convergence of geometric series and the above result imply that

$$\sum_{k \notin \{k_i\}} \Delta_k \leq \frac{1}{1-\gamma_1} \sum_{i=1}^{\infty} \Delta_{k_i} \quad \text{and} \quad \sum_{k=1}^{\infty} \Delta_k \leq \left(1 + \frac{1}{1-\gamma_1} \right) \sum_{i=1}^{\infty} \Delta_{k_i} \leq \infty.$$

This proves the desired result. □

Lemma 5.2.4 will be used to arrive at a contradiction. To obtain this contradiction, I first must show that for k sufficiently large, Algorithm 5.2.1 produces a successful step.

Lemma 5.2.5 *Suppose there exists $\epsilon > 0$ such that $\|\nabla m_k(0)\|_{\mathcal{Z}} \geq \epsilon$ for k sufficiently large. Then, under Assumptions 5.2.2, the ratios, $\{\rho_k\}$, converge to one.*

Proof: By Taylor's theorem, there exists θ_k and η_k on the line segment between $s = 0$ and $s = s_k$ such that

$$\begin{aligned} \text{ared}_k &= \langle \nabla \widehat{J}(z_k), s_k \rangle_{\mathcal{Z}} + \frac{1}{2} \langle \nabla^2 \widehat{J}(\theta_k) s_k, s_k \rangle_{\mathcal{Z}} \\ \text{pred}_k &= \langle \nabla m_k(0), s_k \rangle_{\mathcal{Z}} + \frac{1}{2} \langle \nabla^2 m_k(\eta_k) s_k, s_k \rangle_{\mathcal{Z}}. \end{aligned}$$

These expansions and Assumptions 5.2.2 imply

$$|\text{ared}_k - \text{pred}_k| \leq \xi \Delta_{k-1} \Delta_k + \frac{1}{2}(\kappa_1 + \kappa_2 - 1)\Delta_k^2.$$

Furthermore, the fraction of Cauchy decrease condition, (5.1.2), and the assumption that $\|\nabla m_k(0)\|_{\mathcal{Z}} \geq \epsilon$ imply that for sufficiently large k,

$$\text{pred}_k \geq \kappa_0 \|\nabla m_k(0)\|_{\mathcal{Z}} \min \left\{ \Delta_k, \frac{\|\nabla m_k(0)\|_{\mathcal{Z}}}{\beta_k} \right\} \geq \kappa_0 \epsilon \Delta_k.$$

Combining these inequalities gives

$$|\rho_k - 1| \leq \epsilon_k = \frac{\xi \Delta_{k-1} + \frac{1}{2}(\kappa_1 + \kappa_2 - 1)\Delta_k}{\kappa_0 \epsilon}$$

for sufficiently large k. The sequence $\{\epsilon_k\}$ converges to zero by Lemma 5.2.4, therefore proving the result. □

In addition to achieving a successful step, I must also prove that Algorithm 5.2.1 increases the trust region radius. This result is proved in a similar fashion to Lemma 5.2.5.

Lemma 5.2.6 *Suppose there exists $\epsilon > 0$ such that $\|\nabla m_k(0)\|_{\mathcal{Z}} \geq \epsilon$ for k sufficiently large. Then, under Assumptions 5.2.2, the ratios, $\{\tilde{\rho}_{k+1}\}$, converge to one.*

Proof: By Taylor's theorem, there exists η_{k+1} on the line segment between $s = 0$ and $s = -s_k$ such that

$$m_{k+1}(-s_k) - m_{k+1}(0) = -\langle \nabla m_{k+1}(0), s_k \rangle_{\mathcal{Z}} + \frac{1}{2}\langle \nabla^2 m_{k+1}(\eta_{k+1}) s_k, s_k \rangle_{\mathcal{Z}}.$$

This equality, the expansion of pred_k in the proof of Lemma 5.2.5, Assumptions 5.2.2, and the reverse triangle inequality imply

$$|\text{pred}_k| - |(m_{k+1}(-s_k) - m_{k+1}(0))| \leq \|\nabla m_{k+1}(0) - \nabla m_k(0)\|_{\mathcal{Z}} \Delta_k + (\kappa_2 - 1)\Delta_k^2.$$

To bound this further, notice that

$$\begin{aligned}\|\nabla m_{k+1}(0) - \nabla m_k(0)\|_{\mathcal{Z}} \leq & \|\nabla m_{k+1}(0) - \nabla \widehat{J}(z_k + s_k)\|_{\mathcal{Z}} + \|\nabla \widehat{J}(z_k + s_k) - \nabla \widehat{J}(z_k)\|_{\mathcal{Z}} \\ & + \|\nabla \widehat{J}(z_k) - \nabla m_k(0)\|_{\mathcal{Z}}. \end{aligned} \tag{5.2.2}$$

The first and third expressions on the right hand side of (5.2.2) are bounded using (5.2.1), and the second expression is bounded using the differentiability of $\widehat{J}$, i.e.

$$\|\nabla \widehat{J}(z_k + s_k) - \nabla \widehat{J}(z_k)\|_{\mathcal{Z}} = \left\| \int_0^1 \nabla^2 \widehat{J}(z_k + t s_k) s_k \mathrm{d}t \right\|_{\mathcal{Z}} \leq \kappa_1 \Delta_k.$$

This proves that

$$|\text{pred}_k| - |m_{k+1}(-s_k) - m_{k+1}(0)| \leq (\xi\Delta_k + \xi\Delta_{k-1} + \kappa_1\Delta_k)\Delta_k + (\kappa_2 - 1)\Delta_k^2,$$

which implies the lower bound

$$|m_{k+1}(-s_k) - m_{k+1}(0)| \geq |m_k(s_k) - m_k(0)| - \tilde{\epsilon}_k \Delta_k$$

with $\tilde{\epsilon}_k = (\xi\Delta_k + \xi\Delta_{k-1} + \kappa_1\Delta_k + (\kappa_2 - 1)\Delta_k)$. The fraction of Cauchy decrease condition and the assumption that $\|\nabla m_k(0)\|_{\mathcal{Z}} \geq \epsilon$ imply

$$|m_{k+1}(-s_k) - m_{k+1}(0)| \geq (\kappa_0\epsilon - \tilde{\epsilon}_k)\Delta_k.$$

Since $\tilde{\epsilon}_k$ converges to zero by Lemma 5.2.4, the right hand side of the above inequality is non-negative for sufficiently large k. Following the proof of Lemma 5.2.5, these bounds imply

$$\tilde{\rho}_{k+1} - 1 \leq \frac{\Delta_k(\xi + \frac{1}{2}(\kappa_1 + \kappa_2 - 1))}{\kappa_0\epsilon - \tilde{\epsilon_k}} \to 0 \quad \text{as} \quad k \to \infty.$$

This proves the desired result. □

Combining these results gives the desired result of this section; namely, these results lead to a proof that Algorithm 5.2.1 converges to a first order critical point.

Theorem 5.2.7 *Suppose Assumptions 5.2.2 hold, then*

$$\liminf_{k\to\infty} \|\nabla m_k(0)\|_{\mathcal{Z}} = \liminf_{k\to\infty} \|\nabla \widehat{J}(z_k)\|_{\mathcal{Z}} = 0.$$

Proof: For contradiction, suppose there exists $\epsilon > 0$ such that $\|\nabla m_k(0)\|_{\mathcal{Z}} \geq \epsilon$. By Lemma 5.2.5, for k sufficiently large, there is a successful step, s_k since ρ_k converges to one. By Lemma 5.2.6, for k sufficiently large, the trust region radius must be increased since $\tilde{\rho}_{k+1}$ converges to one. This fact contradicts the result of Lemma 5.2.4. □

5.3 A Framework for Model Adaptivity

Algorithms 5.1.1 and 5.2.1 along with gradient conditions (5.1.3) and (5.2.1), respectively, give natural frameworks for model adaptivity. As long as the modeled gradients satisfy the gradient error bounds, (5.1.3) and (5.2.1), Algorithms 5.1.1 and 5.2.1, respectively, are globally convergent. Furthermore, in many cases it is impossible (or at least computationally infeasible) to compute $\nabla \widehat{J}(z)$. For such problems, the bounds (5.1.3) and (5.2.1) ensure global convergence with out the necessity of computing $\nabla \widehat{J}(z)$ as long as there exists an *a posteriori* error indicator, η, such that

$$\|\nabla \widehat{J}(z_k) - \nabla m_k(0)\|_{\mathcal{Z}} \leq C\eta.$$

Such error indicators exist when the model $m_k(s)$ corresponds to a Galerkin approximation for PDE discretization or a stochastic Galerkin approximation for uncertainty quantification. These indicators come in many forms such as residual based and adjoint based error indicators. When $m_k(s)$ denotes a stochastic collocation approximation (i.e. $m_k(s) = \sigma(\mathcal{L}_{\mathcal{I}_k} j(u(z), z))$ for some tensor product interpolation operator,

$\mathcal{L}_{\mathcal{I}_k}$, built on an admissible index set $\mathcal{I}_k$) I approximate the error indicator, η, as the contribution to the gradient of $m_k(s)$ associated with the indices on the margin of the index set, $\mathcal{M}(\mathcal{I}_k)$. This can be accomplished by employing Algorithm 4.2.3, in which case the margin is denoted as the active set, $\mathcal{A}$, and the error indicator is

$$\eta := \sum_{\mathbf{i} \in \mathcal{A}} \eta_{\mathbf{i}}.$$

In general, this error indicator is heuristic, but seems to work well in practice. Although convergence of the trust region algorithms cannot be guaranteed in any rigorous manner, Corollary 4.2.7 suggests that this choice of error indicator may, in fact, be sufficient.

Now, for simplicity, consider the test problem presented in Section 2.1 with risk measure $\sigma(Y) = E[Y]$ and Algorithm 5.2.1. Denote the objective function based on the index set $\mathcal{I}_k$ as $\widehat{J}_k(z)$ and define the collocation approximate model centered around the iterate z_k as $m_k(s) = \widehat{J}_k(z_k + s)$. The model gradient is

$$\begin{aligned} \nabla m_k(s) = \nabla \widehat{J}_k(z) &= \alpha z + \mathbf{R}^{-1} E \circ \mathcal{L}_{\mathcal{I}_k}[\mathbf{B}^* p(z)] \\ &= \alpha z + \mathbf{R}^{-1} \sum_{\mathbf{i} \in \mathcal{I}_k} E \circ \Delta_{\mathbf{i}} \mathbf{B}^* p(z_k) \end{aligned}$$

where $p(z) = p$ solves the adjoint equation (2.3.2). The loop for satisfying the bound (5.2.1) is presented in Algorithm 5.3.1.

Algorithm 5.3.1 - Model Adaptation:

Let $\mathcal{I}_k = \mathcal{I}_{k-1}$ *and* $\mathcal{A} = \mathcal{M}(\mathcal{I}_k)$

Set $\eta_{\mathbf{i}} = \|E \circ \Delta_{\mathbf{i}} \mathbf{B}^* p(z_k)\|_{\mathcal{Z}}$ *and* $\eta = \sum_{\mathbf{i} \in \mathcal{A}} \eta_{\mathbf{i}}$

Compute $\nabla m_k(0) = \alpha z_k + \mathbf{R}^{-1} \sum_{\mathbf{i} \in \mathcal{I}_k} E \circ \Delta_{\mathbf{i}} \mathbf{B}^* p(z_k)$

Define $TOL = \xi \min \left\{ \|\nabla m_k(0)\|_{\mathcal{Z}}, \Delta_{k-1} \right\}$

while $\eta > TOL$ ***do***

Select $\mathbf{i} \in \mathcal{A}$ *corresponding to the largest* $\eta_{\mathbf{i}}$

Set $\mathcal{A} \leftarrow \mathcal{A} \setminus \{\mathbf{i}\}$ *and* $\mathcal{I}_k \leftarrow \mathcal{I}_k \cup \{\mathbf{i}\}$

Update the error indicator $\eta = \eta - \eta_{\mathbf{i}}$

for $\ell = 1, \ldots, d$ ***do***

Set $\mathbf{j} = \mathbf{i} + \mathbf{e}_\ell$

if $\mathcal{I}_k \cup \{\mathbf{j}\}$ *is admissible* ***then***

Set $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathbf{j}\}$

Set $\eta_{\mathbf{j}} = \|E \circ \Delta_{\mathbf{j}} \mathbf{B}^* p(z_k)\|_{\mathcal{Z}}$

Update the gradient $\nabla m_k(0) \leftarrow \nabla m_k(0) + \mathbf{R}^{-1} E \circ \Delta_{\mathbf{j}} \mathbf{B}^* p(z_k)$

Update the error indicator $\eta \leftarrow \eta + \eta_{\mathbf{j}}$

Update the tolerance $TOL = \xi \min \left\{ \|\nabla m_k(0)\|_{\mathcal{Z}}, \Delta_{k-1} \right\}$

end if

end for

end while

Algorithm 5.3.1 describes the application of adaptive sparse grid stochastic collocation. In particular, Algorithm 5.3.1 employs dimension adaptive sparse grids. Other forms of sparse grid adaptivity exist and can be similarly applied. Some other forms of adaptive sparse grids are domain adaptive sparse grids which refer to partitioning the domain and placing sparse grids in each sub-domain [1] and locally adaptive sparse grids which refers to the use of continuous piecewise linear interpolation as a basis for the sparse grids. In case of local adaptation, one can add points locally and maintain the desirable properties of sparse grids [73].

In general, there may be multiple discretizations necessary to approximately solve the state equation (2.2.4). In the case that $e(u, z; y)$ denotes a nonlinear steady PDE, the stochastic collocation results in Q nonlinear PDEs to be solved, i.e. $e(u_k, z; y_k) = 0$. The finite element method (FEM) is a common approach to discretizing $e(u_k, z; y_k)$ in space. Let $\mathcal{V}_h := \text{span}\{\phi_1, \ldots, \phi_{N_1}\} \subset \mathcal{V}$ be a finite dimensional subspace of

the deterministic state space $\mathcal{V}$ and let $\mathcal{W}_h^* := \text{span}\{\psi_1, \ldots, \psi_{N_2}\} \subset \mathcal{W}^*$ be a finite dimensional subspace of $\mathcal{W}^*$. The Petrov-Galerkin finite element discretization of the k^{th} state equation is

$$\left\langle \psi_m, e\Big(\sum_{n=1}^{N_1} u_{k,n}\phi_n, z; y_k\Big) \right\rangle_{\mathcal{W}^*,\mathcal{W}} = 0 \quad \forall\, m = 1, \ldots, N_2 \tag{5.3.1}$$

and one solves for the vector $\vec{u}_k := (u_{k,1}, \ldots, u_{k,N_1})^\top \in \mathbb{R}^{N_1}$. When coupling FEM with stochastic collocation, there is a natural error splitting. Let $u_{Qh} = u_{Qh}(z) \in L^2_\rho(\Gamma; \mathcal{V}_h)$ denote the stochastic collocation solution computed by solving (5.3.1), then the error in the solution to the state equation can be bounded by

$$\|u(z) - u_{Qh}(z)\|_{\mathcal{U}} \leq \|u(z) - u_Q(z)\|_{\mathcal{U}} + \|u_Q(z) - u_{Qh}(z)\|_{\mathcal{U}}.$$

The first term of this bound is controlled by interpolation, while the second term is controlled by controlling the finite element error. Many approaches exist to adaptively control the error associated with FEM. One popular method is to use residual based error indicators. These indicators are computed using the residual

$$R_k = e\Big(\sum_{n=1}^{N_1} u_{k,n}\phi_n, z; y_k\Big)$$

[35]. Other possible methods are averaging techniques [33] and adjoint based, goal oriented error indicators [61]. Moreover, the authors of [127] employ these FEM error indicators in the context of PDE constrained optimization using the trust region framework.

Similar observations and error control can be performed for the solution to the adjoint equation. Recall that the gradient conditions (5.1.3) and (5.2.1), and the specific form of the gradient place much emphasis on controlling the error in the adjoint state. Care must be taken when controlling the adjoint error because the adjoint directly depends on the solution to the state equation. When solely considering the stochastic collocation discretization, the state error is automatically controlled due to the interpolation properties of $\mathcal{L}_\mathcal{I}$, i.e. $p_k = p_k(u_k(z))$ depends on $u_k = u_k(z)$ which is

exactly the solution to $e(u_k, z; y_k) = 0$ for $k = 1, \ldots, Q$. When considering other discretization techniques, the adjoint error control must also contain state error control. Typically, error bounds for the FEM discretization are

$$\begin{aligned}\|p(u(z)) - p_h(u_h(z))\|_{\mathcal{W}^*} \leq & C_1\|u(z) - u_h(z)\|_{\mathcal{U}} \\ & + C_2\|\widehat{p}(u_h(z) - p_h(u_h(z))\|_{\mathcal{W}^*}\end{aligned}$$

where $p = p(u(z)) \in \mathcal{W}^*$ is the solution to the infinite dimensional adjoint equation (2.3.2) and $\widehat{p} = \widehat{p}(u_h(z)) \in \mathcal{W}^*$ is the solution to the infinite dimensional adjoint equation (2.3.2) with $u(z)$ replaced by $u_h(z)$.

With the above discussion in mind, one can incorporate both the finite element and collocation error in the adaptive loop 5.3.1 by performing the error splitting

$$\|\nabla\widehat{J}(z) - \nabla\widehat{J}_{Qh}(z)\|_{\mathcal{Z}} \leq \|\nabla\widehat{J}(z) - \nabla\widehat{J}_{Q}(z)\|_{\mathcal{Z}} + \|\nabla\widehat{J}_{Q}(z) - \nabla\widehat{J}_{Qh}(z)\|_{\mathcal{Z}}$$

where $\widehat{J}_Q(z)$ denotes the collocation discretization of the objective function and $\widehat{J}_{Qh}(z)$ denotes the collocation and finite element discretized objective function. Note, the first term only deals with collocation error and the second term only deals with finite element error. Now, suppose one is considering adding an index, $\mathbf{k}$, to the current generalized sparse grid index set, $\mathcal{I}$. If $\mathcal{I} \cup \{\mathbf{k}\}$ remains admissible, then one can derive the error indicator $\eta_{\mathbf{k}}$ (again, note that this is only collocation discretized and, thus, not necessarily computable). If one defines a similar error indicator for the stochastic collocation finite element discretized problem, $\eta_{\mathbf{k}}^h$, then one can approximate $\|\eta_{\mathbf{k}}\|_{\mathcal{Z}}$ as

$$\|\eta_{\mathbf{k}}\|_{\mathcal{Z}} \leq \|\eta_{\mathbf{k}} - \eta_{\mathbf{k}}^h\|_{\mathcal{Z}} + \|\eta_{\mathbf{k}}^h\|_{\mathcal{Z}}.$$

Here, the first term is controlled by the finite element error. Controlling the error in the first term so that

$$\|\eta_{\mathbf{k}} - \eta_{\mathbf{k}}^h\|_{\mathcal{Z}} \leq c\|\eta_{\mathbf{k}}^h\|_{\mathcal{Z}},$$

where $c > 0$ is fixed, will give the error bound

$$\|\eta_{\mathbf{k}}\|_{\mathcal{Z}} \leq (c+1)\|\eta_{\mathbf{k}}^h\|_{\mathcal{Z}}$$

which depends only on the computable quantity, $\|\eta_{\mathbf{k}}^h\|_{\mathcal{Z}}$.

In the case of time dependent PDE operators $e(u, z; y)$, one can consider adaptive time stepping and adaptive basis selection for model order reduction. In [77] and [78, 79], the authors perform adaptive time stepping for the optimal control of deterministic parabolic PDEs. In these works, the authors incorporate error control for both finite element and time stepping discretization. On the other hand, the authors of [48] employ the trust region framework to adaptive build reduced order models for parabolic control problems. In this work, the reduced order models are built using proper orthogonal decomposition. In general, the trust region algorithms are flexible enough to handle any sort of model adaptivity so long as the gradient conditions (5.1.3) and (5.2.1) are satisfied. Hence, it is possible to incorporate these additional adaptive strategies using Algorithm 5.3.1 and similar arguments as above.

Chapter 6

Implementation Details

This chapter is dedicated to considerations for a high performance implementation of the adaptive stochastic collocation and trust region framework described in this thesis. When discretized, the optimization problems discussed here are extremely high dimensional and computationally intensive. The numerical solution of these problems is challenging to say the least. When implementing the methods discussed in this thesis, one must exploit the natural parallelism of the stochastic collocation method. Furthermore, one must use efficient large-scale nonlinear programming techniques in the implementation of Algorithms 5.1.1 and 5.2.1. I will first discuss implementation of a high fidelity discretization of the objective function and ways to exploit multiple forms of parallelism in function evaluations and derivative computations. I will then present nonlinear programming ideas to accelerate the trust region approach.

6.1 High Fidelity Objective Computation

The implementation of Algorithms 5.1.1, 5.2.1, and 5.3.1 require careful consideration. There are many implementation details that can significantly improve the performance of these algorithms. When using either trust region algorithm, one must compute the ratio between actual and predicted decrease, ρ_k. Each evaluation of this

ratio requires the computation of the objective function, $\widehat{J}(z)$. In general, the objective function need not be discretized if one is able to evaluate the infinite dimensional function $\widehat{J}(z)$. This is almost never the case. Since this thesis is concerned with discretization of the stochastic space, I will discretize $\widehat{J}(z)$ using stochastic collocation. This discretization must be high fidelity in order for adaptivity to be meaningful. Thus, for this high fidelity discretization, I will use isotropic Smolyak sparse grids of high level (c.f. see the definition in Remark 4.2.2). Recall here that Corollary 3.4.2 and Theorem 3.4.4 prove convergence for such discretizations. Using isotropic Smolyak sparse grids for the high fidelity discretization is advantageous because these interpolation operators require much fewer interpolation knots as opposed to full tensor product interpolation. Furthermore, *a priori* knowledge of the anisotropy in the optimization problem is typically not available. Without this knowledge, one cannot build meaningful anisotropic sparse grids. As a consequence of the adaptive loop (5.3.1), the resulting adapted sparse grid contains information concerning the anisotropy associated with the optimization problem. This anisotropy can be visualized using the final index set, $\mathcal{I}$, and can be used to help characterize the importance of the random variables in Γ.

Although isotropic Smolyak sparse grids provide a reduction in computational cost when compared to full tensor product grids, they still pose a possibly enormous computational burden. Adaptive collocation allows for relatively cheap step computation in the trust region framework, but this may be outweighed by the high fidelity function evaluations required at each iteration. Since my adaptive framework uses low order adapted sparse grids to compute gradients and Hessian information, there is a clear advantage when compared to Newton-CG applied to the high fidelity problem as long as many trust region iterations are not required. Note that in the adaptive framework, if the knots associated with the adapted sparse grids are subsets of those associated with the high fidelity sparse grid (i.e. nested knots), one can re-use state computations. Further efficiency can be achieved by exploiting the almost trivial

parallelism of the stochastic collocation method.

6.2 Parallel Collocation and Linear Algebra

The stochastic collocation method described in this thesis lends itself naturally to parallel implementation. The stochastic collocation method for the solution of (2.2.4) requires the decoupled solutions of

$$\tilde{e}(u_k, z; y_k) = 0 \quad \forall\, k = 1, \ldots, Q$$

for fixed $z \in \mathcal{Z}$. Additionally, in the optimization context, one must solve the adjoint equation (2.3.2)

$$\tilde{e}_u(u_k, z; y_k)^{\lambda}_k + j_v(u_k, z) = 0 \quad \forall\, k = 1, \ldots, Q.$$

The solution to the adjoint equation at the k^{th} collocation point, λ_k, depends on the solution to the state equation at the k^{th} collocation point, u_k. Therefore, the state and adjoint equations must be solved in serial. One can solve the state and adjoint equations at different collocation points concurrently. Now, consider the approximation operator, $\mathcal{L}_Q$, with polynomial representation

$$(\mathcal{L}_Q u)(y) = \sum_{k=1}^{Q} P_k(y) u_k.$$

The parallel function evaluation and gradient computation algorithm is listed in Algorithm 6.2.1.

Algorithm 6.2.1 - Parallel Function Evaluation and Gradient Computation:

Given $z \in \mathcal{Z}$ *and* $\{y_k\}_{k=1}^{Q} \subset \Gamma$*;*

for $k = 1, \ldots, Q$ ***do***

Compute u_k *which solves* $\tilde{e}(u_k, z; y_k) = 0$*;*

Compute $j_k = j(u_k, z)$*;*

Compute λ_k *which solves* $\tilde{e}_u(u_k, z; y_k)^* \lambda_k + j_u(u_k, z) = 0$.

end for

Compute $\widehat{J}_Q(z) = \sigma\Big(\sum_{k=1}^{Q} P_k(y) j_k\Big)$*;*

Compute the derivative of $\widehat{J}_Q(z)$ *as*

$$\widehat{J}'_Q(z) = \sum_{k=1}^{Q} \vartheta_k \Big\{ e_z(u_k(z), z; y_k)^* \lambda_k + j_z(u_k(z), z) \Big\}$$

where $\vartheta_k = E\Big[\sigma'\Big(\sum_{\ell=1}^{Q} P_\ell(y) j(u_\ell(z), z)\Big) P_k(y)\Big]$.

Algorithm 6.2.1 demonstrates that the state and adjoint computations can all be performed in parallel on multiple processors, but each process must communicate to compute objective functions and gradients. The application of the Hessian of $\widehat{J}_Q(z)$ to a vector, $v \in \mathcal{Z}$, exhibits the same structure and parallelism as Algorithm 6.2.1. For more information on the computation of second order information see [67].

The parallelism of the stochastic collocation discretization is essential for high dimensional stochastic spaces, Γ. In general, the total number of collocation points, Q, is extremely large and the only feasible approach to solving these state and adjoint equations is to exploit this parallelism. Aside from the immense dimension of the collocation space, it is often the case that the solution to the state and adjoint equations at each collocation point is also computationally intense. This is the case when $e(u, z; y)$ denotes a partial differential equation. Common PDE discretization techniques often yield extremely high dimensional nonlinear and linear systems to be solved for the state and adjoint equations, respectively. With the existence of many parallel and distributed linear algebra tools, it is common practice to solve the state and adjoint equations using distributed Krylov subspace methods. Furthermore, in

the case of PDE constraints, much work has gone into developing parallel preconditioners for solving the linearized state equation and the adjoint equation. Many such preconditioners exist, such as domain decomposition and multi-level preconditioners. Including distributed linear algebra to Algorithm 6.2.1 gives an efficient, fully parallel approach to computing derivatives. The use of linear algebra and collocation parallelism is not completely straightforward though. Suppose I have a total of nproc processes available for computation and suppose I wish to dedicate nLA processes to linear algebra. Here, I assume nLA divides nproc. I can then split the nproc processes into nSC $= \frac{\text{nproc}}{\text{nLA}}$ groups of nLA processes. Each of these nSC groups is given a subset of the collocation points and nLA processes to perform distributed linear algebra computations. This scheme is depicted in Figure 6.1.

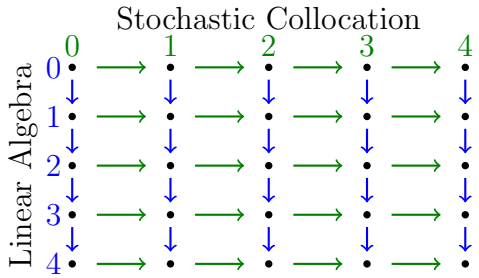


Figure 6.1: Depiction of the communication pattern used to incorporate distributed linear algebra and parallel stochastic collocation computations.

6.3 Nonlinear Programming Considerations

Although, function evaluations and derivative computations can be performed exploiting parallelism in the linear algebra and the stochastic collocation, these computations are still extremely expensive. For this reason, it is crucial to use efficient nonlinear programming techniques in the implementation of Algorithms 5.1.1 and 5.2.1. In the case that $\widehat{J}_Q(z)$ is quadratic (or if $m_k(s)$ are chosen to be quadratic approximations

to $\widehat{J}_Q(z)$), the trust region subproblem (5.1.1) can be solved using the truncated conjugate gradient (CG) method [41]. Truncated CG is a generalization to the standard CG algorithm which exits if the current CG solution is outside the trust region or if the algorithm finds a direction of negative curvature. Truncated CG requires a function that applies the Hessian of m_k to a vector. As mentioned above, Hessian times a vector computations are similar to Algorithm 6.2.1 and require $2Q$ linear PDE solves. Although this can be performed using the parallelism described above, if CG requires many iterations, the trust region approach quickly becomes computationally infeasible. Adaptivity helps in this aspect because for early trust region iterations, the number of collocation points is small and thus, only a few linear PDEs must be solved at each CG iteration.

Another method of reducing the number of CG iterations is to use preconditioning. In [67], the author explores using low order sparse grid approximations to precondition the Hessian. This method works well for small problems, but when the stochastic dimension is large, even low order sparse grids have many collocation points. In this work, I use limited memory quasi-Newton approximations to the inverse of the Hessian to automatically precondition the CG iterations. The use of quasi-Newton preconditioners was presented in [80] in the general context of nonlinear programming and in [23] in the context of flow control. The results presented in [80] and [23] appear to be inconclusive on whether or not quasi-Newton preconditioning is advantageous. For the problems of interest to this thesis, quasi-Newton preconditioning is almost essential for solving large problems.

In the case that Hessian information is unavailable, quasi-Newton approximations can be used to generate quadratic approximations of $\widehat{J}_Q(z)$. In this case, one can use the quasi-Newton approximation of the Hessian and the inverse Hessian to construct a double dog leg approach for the solution of (5.1.1) [66]. This approach approximates the step by constructing a piecewise linear path between the Cauchy point and the quasi-Newton point (double dog leg curve). If the quasi-Newton point is within the

trust region, then the algorithm chooses this as the new step. If the quasi-Newton point is outside of the trust region, then the step is chosen as the point for which the double dogleg curve intersects the trust region. On the other hand, one can also use the quasi-Newton approximation of the Hessian with truncated CG to solve the subproblem (5.1.1). The CG iterations in this approach are inexpensive and the CG algorithm gives better steps when the quasi-Newton point is outside of the trust region.

Chapter 7

Numerical Examples

In this chapter, I will present a variety of numerical examples demonstrating the power and necessity of adaptivity in solving optimization problems governed by PDEs with uncertain coefficients. Throughout this section, I will refer to ten separate optimization algorithms. These algorithms are denoted: NEWTONCG, NEWTONCG + BFGS, TRCG HESS, TRCG HESS + BFGS, TRCG BFGS, TRDOGLEG BFGS, RTRCG HESS, RTRCG HESS + BFGS, RTRCG BFGS, and RTRDOGLEG BFGS. NEWTONCG denotes Newton-CG using a high fidelity stochastic collocation discretization of the optimization problem. NEWTONCG + BFGS is Newton-CG using limited memory BFGS preconditioning. TRCG stands for trust region and RTRCG stands for retrospective trust region where the trust region subproblems are solved using truncated CG. The suffix "HESS" refers to using Hessian information and the addition of "+ BFGS" denotes the use of limited memory BFGS preconditioning. The suffix "BFGS" denotes the use of limited memory BFGS to approximate the Hessian. Finally, TRDOGLEG BFGS and RTRDOGLEG BFGS refer to the use of the double dogleg approach combined with limited memory BFGS Hessian approximations to approximately solve the trust region subproblem.

7.1 One Dimensional Optimal Control

The one dimensional examples in this section all correspond to the distributed control of the steady heat equation. In this section, I will present three examples each demonstrating different forms of randomness. One particularly nice feature of my adaptive approach is that Algorithm 5.3.1 exploits anisotropy in the stochastic dependence of the state and adjoint equations (i.e. dependence on the different directions Γ_k). This anisotropy ultimately can reduce the number of PDE solves required for the computation of derivative information.

Let the physical domain be the bounded interval, $D \subset \mathbb{R}$, and let $\Gamma \subset \mathbb{R}^M$ denote the stochastic image space. Γ is endowed with the uniform probability density $\rho(y)$. Throughout this section, I will consider the quadratic control problem

$$\min_{z \in \mathcal{Z}} \widehat{J}(z) := \frac{1}{2} E\Big[\|u(z) - \bar{v}\|^2_{L^2(D)}\Big] + \frac{\alpha}{2}\|z\|^2_{L^2(D)}$$

where $u(y) = u(y; z) \in H^1_0(D)$ for all $y \in \Gamma$ solves the state equation

$$-\epsilon(y)\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(y,x) = f(y,x) + z(x), \quad (y,x) \in \Gamma \times D \tag{7.1.1}$$

$$u(y,x) = 0, \quad (y,x) \in \Gamma \times \partial D$$

To relate this to the test problem in Section 2.1, $\mathcal{Z} = L^2(D)$, $\mathcal{V} = H^1_0(D)$, $\mathcal{W} = H^{-1}(D)$, $\mathcal{H} = L^2(D)$, and $\mathbf{Q}$ is defined by

$$\langle \mathbf{Q}u, v\rangle_{H^{-1}(D), H^1_0(D)} = \int_D u(x)v(x)\mathrm{d}x.$$

The first two examples will use truncated KL expansion diffusivity coefficients. The isotropic nature of the solution $u(y) = u(y, z)$ of (7.1.1) is characterized by the decay of the eigenvalues of the covariance function [106]. This decay is not sufficient to characterize the dependence of the optimization problem on the directions Γ_k. For simplicity, I will consider the state operator, $\widehat{\mathbf{A}}(y) \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ for $y \in \Gamma$, given as a truncated KL expansion, i.e.

$$\widehat{\mathbf{A}}(y) = \mathbf{A}_0 + \sum_{k=1}^{M} \sqrt{\lambda_k}\mathbf{A}_k y_k$$

and the operator form of the state equation (7.1.1)

$$\widehat{\mathbf{A}}(y)u(y) + \widehat{\mathbf{B}}z = 0.$$

Define the deterministic reduced cost functional

$$\widehat{j}(z;y) := j(u(y;z),z) = \frac{1}{2}\|u(y;z) - \bar{v}\|^2_{L^2(D)} + \frac{\alpha}{2}\|z\|^2_{L^2(D)}$$

where $u(y) = u(y;z)$ solves the state equation (7.1.1). Fix $z \in \mathcal{Z}$ and define the function $f(y) := j(z;y)$. One can write the second order Taylor approximation of $f(y)$ centered around $\bar{y} = E[y]$ as

$$f(y) = f(\bar{y}) + \nabla f(\bar{y})^\top (y-\bar{y}) + \frac{1}{2}(y-\bar{y})^\top \nabla^2 f(y + t(\bar{y}-y))(y-\bar{y}) \quad \text{for } t \in (0,1).$$

Note that since $\nabla f(\bar{y})$ is independent of $y \in \Gamma$, $E[\nabla f(\bar{y})^\top (y-\bar{y})] = 0$. Now, one can explicitly write down the Hessian $\nabla^2 f(\zeta)$ as

$$\nabla^2 f(\zeta) = (\partial^2_{uu} j(u(\zeta;z),z)\partial_y u(\zeta,z)) + \partial_u j(u(\zeta,z),z)(\partial^2_{yy} u(\zeta,z)).$$

The derivatives, $\partial_y u(\zeta,z)$ and $\partial^2_{yy} u(\zeta,z)$, can be computed via implicitly differentiating (7.1.1). The first partial derivative of u with respect to y_k solves

$$\widehat{\mathbf{A}}(\zeta)\partial_{y_k} u(\zeta,z) + \sqrt{\lambda_k}\mathbf{A}_k u(\zeta,z) = 0$$

and the second partial derivative solves

$$\widehat{\mathbf{A}}(\zeta)\partial^2_{y_k y_j} u(\zeta,z) + \sqrt{\lambda_j}\mathbf{A}_j \partial_{y_k} u(\zeta,z) + \sqrt{\lambda_k}\mathbf{A}_k \partial_{y_j} u(\zeta,z) = 0.$$

From these derivatives, it is clear to see that the elements of the Hessian matrix, $\nabla^2 f(\zeta)$ satisfy the upper bound

$$|\widehat{J}(z) - j(u(\bar{y};z),z)| = |E[f(y)] - f(\bar{y})| \leq C \sum_{j=1}^{M}\sum_{k=1}^{M} \sqrt{\lambda_j \lambda_k}.$$

This simple analysis demonstrates the fact that even if the operator $\mathbf{A}(y)$ is anisotropic with respect to the dimensions Γ_k for $k = 1, \ldots, M$, the objective function may depend isotropically on the parameters $y \in \Gamma$.

Remark 7.1.1 *When considering the state equation*

$$\widehat{\mathbf{A}}u(y) + \widehat{\mathbf{B}}z = \mathbf{b}(y),$$

where $\widehat{\mathbf{A}}$ *is deterministic and* $\mathbf{b}(y)$ *is given as a truncated KL expansion, a similar analysis can be performed. In this case, the solution* $u(y) = u(y, z) \in \mathcal{V}$ *for all* $y \in \Gamma$ *depends linearly on* $y \in \Gamma$*. Thus, plugging* $u(y)$ *into the quadratic objective function,* $j(u(y, z), z)$*, gives a quadratic dependence of* $f(y) = \widehat{j}(z; y)$ *on the stochastic variables* $y \in \Gamma$*. Since this dependence is quadratic,* $f(y)$ *is exactly equal to its second order Taylor polynomial and the Hessian is constant with respect to* $y \in \Gamma$*. This gives the upper bound*

$$|\widehat{J}(z) - j(u(\bar{y}; z), z)| = |E[f(y) - f(\bar{y})| \leq C \sum_{k=1}^{M} \lambda_k.$$

Hence, the decay in the eigenvalues, λ_k*, completely controls the anisotropic dependence of the objective function on* $y \in \Gamma$*.*

7.1.1 An Isotropic Example

This example is presented in [28] and serves as an example that does not result in an anisotropic sparse grid. The physical domain is $D = (-1, 1)$ and the diffusivity coefficients are defined as the truncated KL expansion

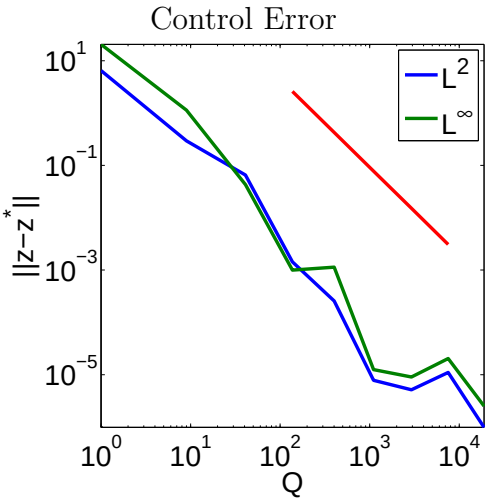
$$\epsilon(y, x) = \epsilon_0(x) + \sum_{k=1}^{M} \frac{1}{4} \epsilon_k(x) y_k$$

where $\epsilon_0(x) \equiv 2$ and $\epsilon_k(x)$ are the L^2 normalized Lagrange polynomials built Gauss-Legendre interpolation knots. The random variables, y_k, are assumed to be uniformly distributed on $\Gamma = [-1, 1]^M$. The regularization parameter is $\alpha = 10^{-6}$ and the target profile is

$$\bar{v}(x) = 2 + \text{sign}(\cos(\pi x)).$$

For the numerical examples, I fix $M = 4$ terms in the KL expansion. I have discretized this optimal control problem in space using continuous piecewise linear

FEM on a uniform mesh of 128 intervals. Figure 7.1 depicts the stochastic collocation error in the optimal. For the "true" solution, I solve the optimal control problem using a level eleven isotropic Smolyak sparse grid ($Q = 271617$). Here, the red line has slope $\nu = 1.7$. This slope was computed using least squares. The table in Figure 7.1 contains the $L^2(D)$ error associated with different levels of isotropic Smolyak sparse grids (ℓ) and the corresponding number of sparse grid cubature knots (Q). In this table, "rate" refers to the slope between two consecutive points.



ℓ	Q	Error	Rate
0	1	6.48111191e+00	
1	9	2.93765793e-01	1.408
2	41	6.53655636e-01	0.991
3	137	1.43160490e-03	3.167
4	401	2.55996314e-04	1.603
5	1105	7.80872629e-06	3.443
6	2925	5.13762335e-06	0.430
7	7537	1.10353993e-05	-0.808
8	18945	9.86769644e-07	2.619

Figure 7.1: (Left) Collocation error in the optimal controls. The red line denotes the least squares fit for the collocation. The estimated convergence rate is $\nu = 1.7$. (Right) L^2 error and associated rate of decrease for the optimal controls.

To use the adaptive collocation and trust region framework described in this thesis, I employ level five isotropic Smolyak sparse grids built on one dimensional Clenshaw-Curtis interpolation knots ($Q = 1105$) for the high fidelity approximation of the objective function. In Figure 7.2, I have plotted the optimal control on the left and the expected value of the solution to the state equation computed using the optimal control. In addition, I have added one and two standard deviation intervals around the expected value of the state. Furthermore, in Table 7.1, I compare ten different algorithms described above.

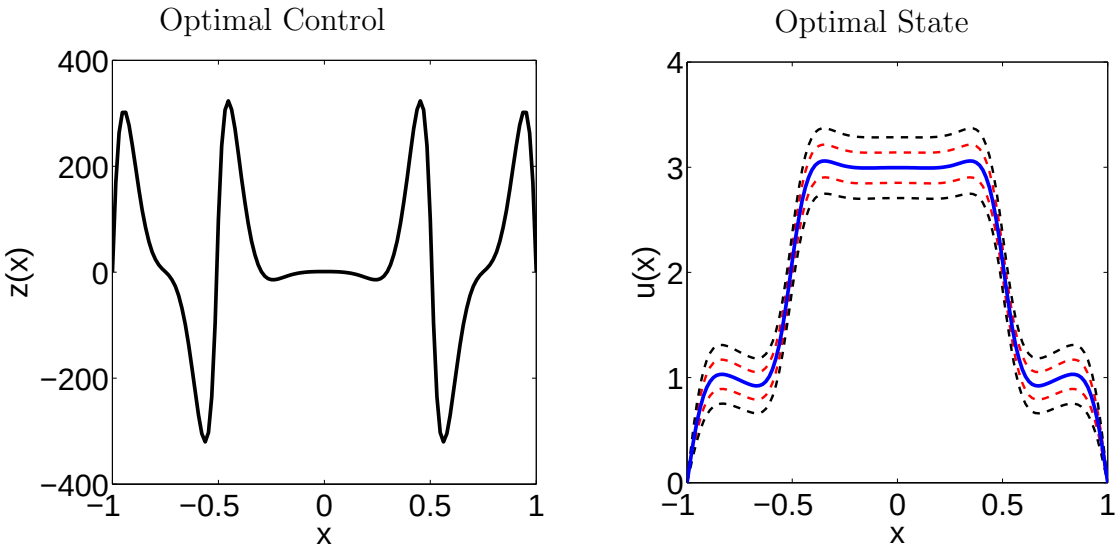


Figure 7.2: (Left) Computed optimal control. (Right) Expected value of optimal state (blue solid line) plus one (red dashed line) and two (black dashed line) standard deviations.

	TR	Adaptive	PDE	CP	Reduction
NEWTONCG	7	0	224,315	1,105	
NEWTONCG + BFGS	6	0	114,920	1,105	1.95
TRCG HESS	6	88	110,367	769	2.03
TRCG HESS + BFGS	5	89	46,934	809	4.78
TRCG BFGS	60	71	196,671	721	1.14
TRDOGLEG BFGS	61	71	202,825	721	1.11
RTRCG HESS	6	88	113,561	769	1.98
RTRCG HESS + BFGS	5	89	48,630	809	4.61
RTRCG BFGS	57	88	240,162	769	0.93
RTRDOGLEG BFGS	59	71	234,543	721	0.96

Table 7.1: This table contains the total number of outer iterations (TR), the total number of adaptive steps (Adaptive), the total number of PDE solves (PDE), the total number of collocation points in the final sparse grid (CP), and the reduction factor of total PDE solves required by the specified algorithm versus Newton-CG (Reduction).

7.1.2 A Mildly Anisotropic Example

In this example, the physical domain is $D = (0, 1)$. To construct the diffusivity coefficients, define the sets $S_k := (\frac{k-1}{M}, \frac{k}{M}]$ for $k = 1, \ldots, M$. The diffusivity parameter is defined as

$$\epsilon(y, x) = \sum_{k=1}^{M} \left\{ (M - k + 0.01) + \frac{M - k}{k^3} y_k \right\} \chi_{S_k}(x).$$

This type of diffusion parameters is called checkerboard diffusion and is often used for numerical examples. Each interval in the checkerboard diffusion parameter has decreasing importance (i.e. the diffusivity on each interval is scaled by $\frac{M-k}{k^3}$). This type of diffusion was studied in [39]. The stochastic image space is $\Gamma := [0, 1]^M$ endowed with the uniform distribution, $\rho(y) \equiv 1$. The regularization parameter is set to $\alpha = 10^{-6}$ and the desired profile is given by

$$\bar{v}(x) = 2x + \sin(2\pi x).$$

I have discretized this optimal control problem using piecewise linear finite elements on a uniform mesh of 128 intervals. I choose $M = 4$ random variables and, for the high fidelity approximation of the objective function, I use a level five isotropic Smolyak sparse grid built on one dimensional Clenshaw-Curtis interpolation knots ($Q = 1105$). In Table 7.2, I compare the ten different algorithms described above. This table clearly demonstrates the advantage of using my adaptive approach. With no preconditioning, the adaptive approach experiences about a seven fold reduction in the number of PDE solves required to obtain the optimal controls. Adding limited memory BFGS preconditioning further reduces the number of PDE solves required and results in a ten fold reduction. Figure 7.3 depicts the computed optimal control (left) and the expected value of the optimal state (blue solid line) plus one (red dashed line) and two (black dashed line) standard deviation intervals. Figure 7.3 clearly demonstrates the mild anisotropy associated with this optimization problem. One will notice that the standard deviation is smaller for $x > 0$ and decreases as x tends toward one. Figure 7.4 demonstrates the stochastic collocation discretization

error. The red line in the left image is a least squares fit to the error curve. The least squares estimated slope is $\nu = 3.5$. The table on the left contains the $L^2(D)$ error for the different levels of isotropic Smolyak sparse grid and their associated slopes between two consecutive levels. Finally, Figure 7.5 depicts the optimal controls (left) corresponding to the deterministic substitute problem where the random variables $y \in \Gamma$ were replaced with $\bar{y} = E[y]$. The right image in Figure 7.5 depicts the absolute error between the controls computed for the deterministic problem and the controls computed for the stochastic problem.

	TR	Adaptive	PDE	CP	Reduction
NEWTONCG	6	0	143650	1105	
NEWTONCG + BFGS	6	0	90610	1105	1.59
TRCG HESS	6	21	20685	185	6.94
TRCG HESS + BFGS	5	27	14058	261	10.22
TRCG BFGS	41	19	59848	153	2.40
TRDOGLEG BFGS	51	18	75347	141	1.91
RTRCG HESS	6	21	21431	185	6.70
RTRCG HESS + BFGS	5	27	14626	261	9.82
RTRCG BFGS	41	19	64660	153	2.22
RTRDOGLEG BFGS	51	18	81882	141	1.75

Table 7.2: Algorithm comparison for checkerboard diffusivity one dimensional example.

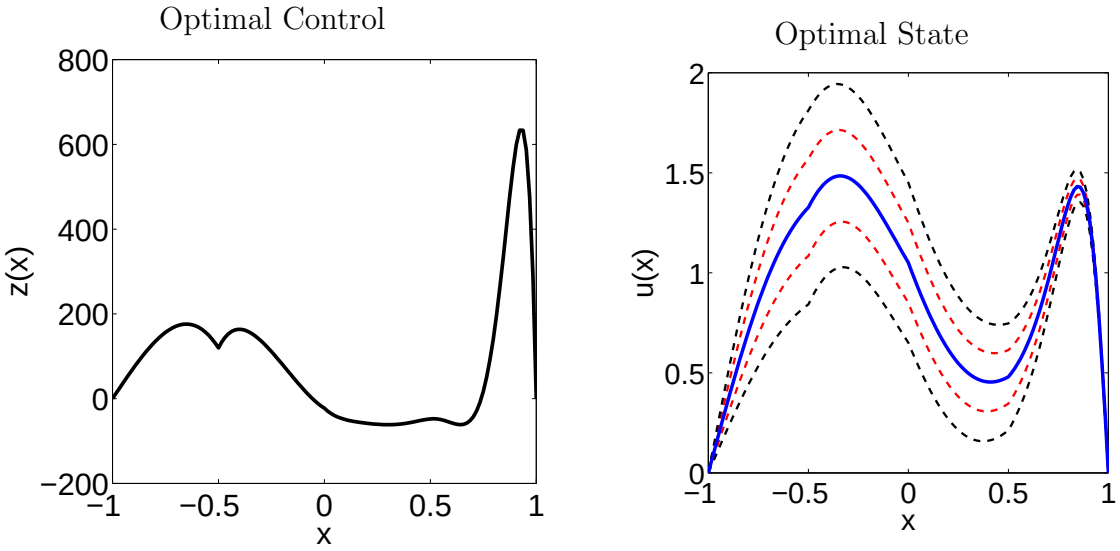
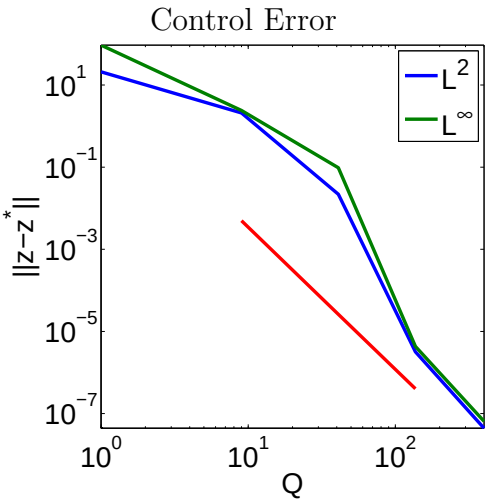


Figure 7.3: (Left) Computed optimal control. (Right) Expected value of optimal state (blue solid line) plus one (red dashed line) and two (black dashed line) standard deviations.



ℓ	Q	Error	Rate
0	1	2.09105168e+01	
1	9	2.09008865e+00	1.048
2	41	2.18699979e-02	3.007
3	137	3.18712385e-06	7.322
4	401	4.41271949e-08	3.985
5	1105	0.0	

Figure 7.4: (Left) Collocation error in the optimal controls. The least squares fit red line has slope $\nu = 3.5$. (Right) L^2 error and associated rate of decrease for the optimal controls.

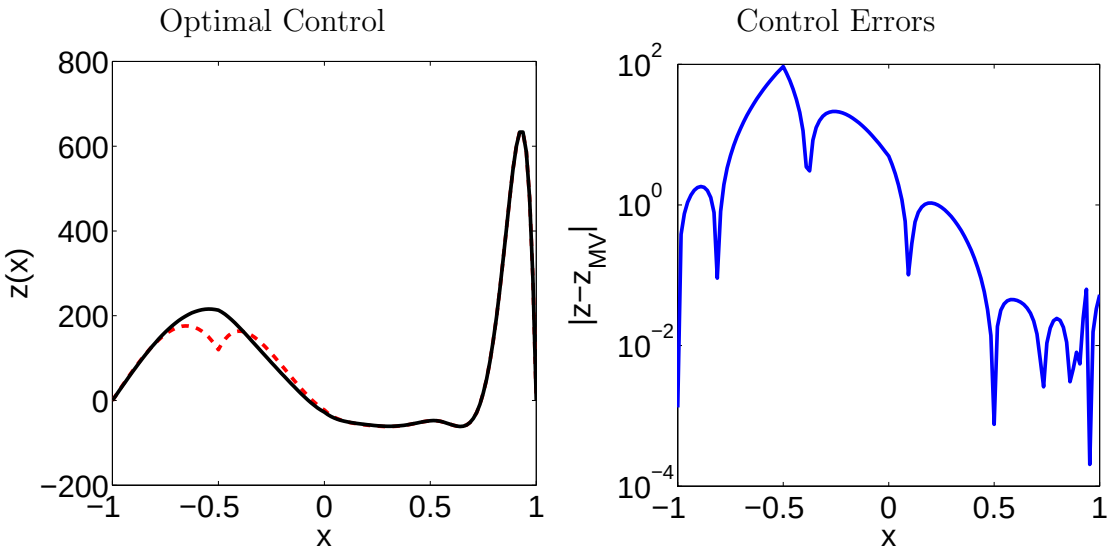


Figure 7.5: (Left) The optimal controls for the deterministic problem with $y \in \Gamma$ replaced by $\bar{y} = E[y]$ (solid black line). The red dashed line is the control computed via the stochastic problem. (Right) Errors between the optimal controls for the stochastic problem and the optimal controls for the mean value problem.

7.1.3 An Anisotropic Example

This example is motivated by subsurface flow control through fractured media. In particular, I am interested in the situation where the location of the fractures is uncertain. Furthermore, this example investigates the effects of discontinuous diffusion parameters on the convergence of the stochastic collocation method. The diffusivity parameters in the example are not given as a truncated KL expansion. Furthermore, this example incorporates a stochastic right hand side. The physical domain is $D = (-1, 1)$ and the stochastic image space is $\Gamma = [-0.1, 0.1] \times [-0.5, 0.5]$ endowed with the uniform probability density. The diffusion parameter is defined as

$$\epsilon(y, x) = \epsilon_1 \chi_{(-1,y_1)}(x) + \epsilon_2 \chi_{(y_1,1)}(x)$$

with $\epsilon_1 = 0.1$, $\epsilon_2 = 10$ and the forcing term is defined as

$$f(y, x) = e^{-(x-y_2)^2}.$$

The regularization parameter is set to $\alpha = 10^{-4}$ and target profile is $\bar{v} \equiv 1$.

I discretized the state equation in space using continuous piecewise linear finite elements on a uniform mesh of $N = 128$ intervals. In order to obtain accurate results, the discontinuity in ϵ should be aligned with mesh vertices. To do this, each sample point, $y_1 \in [-0.1, 0.1]$, can be added as a vertex to the mesh. The collocation space is built on one dimensional Gauss-Patterson interpolation knots. The high fidelity collocation discretization is performed using a level seven isotropic Smolyak sparse grid ($Q = 1793$). The optimization results are depicted in Figure 7.6 and Figure 7.7. Figure 7.6 illustrates the adaptive sparse grid at the final step of optimization. The right image contains the active (red) and old (blue) index sets. The union of these two sets gives the generalized sparse grid index set, $\mathcal{I}$. The left figure depicts the resulting sparse grid of collocation points corresponding to $\mathcal{I}$. From this index set, the anisotropy associated with this optimization problem is clear; that is, much more refinement is necessary in the y_1 direction as opposed to the y_2 direction. Figure 7.7

displays both the optimal controls and the expected value of the state plus one and two standard deviation intervals. The iteration history for the trust region framework is displayed in Table 7.3.

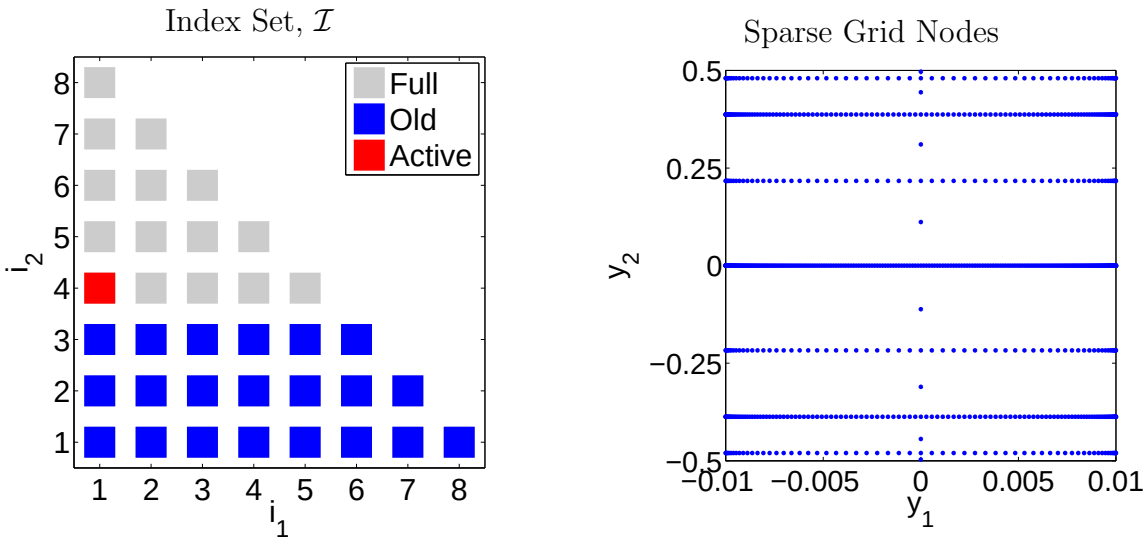


Figure 7.6: (Left) Generalized sparse grid index set. The red blocks denote "active" indices and the blue blocks denote "old" indices. The gray blocks denote the indices in the isotropic Smolyak index set of level eight. (Right) Collocation points corresponding to the index set $\mathcal{I} = \mathcal{A} \cup \mathcal{O}$.

Figure 7.8 depicts the optimal controls (left) corresponding to the deterministic substitute problem where the random variables $y \in \Gamma$ were replaced with $\bar{y} = E[y]$. The right image in Figure 7.8 depicts the absolute error between the controls computed for the deterministic problem and the controls computed for the stochastic problem. Figure 7.9 depicts the stochastic collocation error associated with different levels of the isotropic Smolyak sparse grid. The least squares estimated convergence rate (red line) is $\nu = 0.7$. The convergence rate is severely diminished from the previous one dimensional example. This convergence rate is expected due to the lack of smoothness in the state and adjoint equations with respect to y_1. In fact, this convergence rate is roughly the same as Monte Carlo (i.e. $\nu = 0.5$).

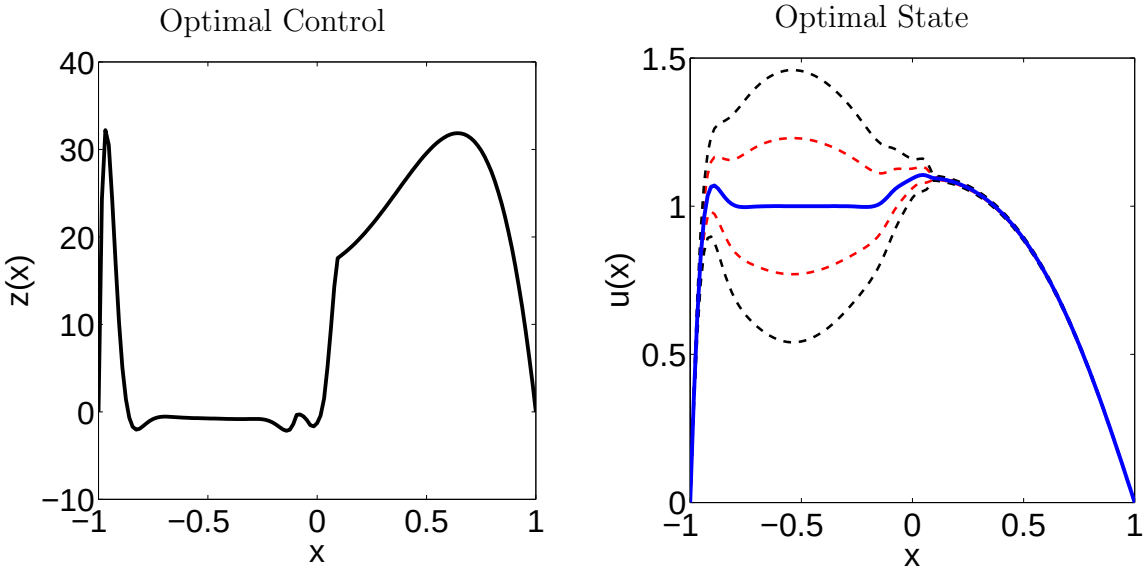


Figure 7.7: (Left) Computed optimal control. (Right) Expected value of computed optimal state with one and two standard deviation intervals added.

k	$\widehat{J}(z_k)$	$\|\nabla \widehat{J}_{\mathcal{I}}(z_k)\|_{\mathcal{Z}}$	$\|s_k\|_{\mathcal{Z}}$	Δ_k	CG	Adaptive	CP
1	2.264427e-01	1.855366e-02	1.528843e+02	1000	5	6	1
2	1.509245e-01	8.555873e-04	4.918603e+01	2500	5	7	49
3	1.349774e-01	6.153557e-05	8.451017e+01	5000	10	2	385
4	1.346243e-01	3.457096e-06	1.502707e+01	5000	12	2	513
5	1.346233e-01	2.368099e-07	2.097707e+00	5000	12	2	545
6	1.346233e-01	1.022377e-08	1.196515e-01	5000	11	2	641
7	1.346233e-01	5.780752e-10	8.693302e-03	5000	13	0	769

Table 7.3: **(Iteration History)** k is the number of trust region iteration, $\widehat{J}(z_k)$ is the objective function value, $\|\nabla \widehat{J}_{\mathcal{I}}(z_k)\|_{\mathcal{Z}}$ is the model gradient norm value, $\|s_k\|_{\mathcal{Z}}$ is the step size, Δ_k is the trust region radius, CG is the number of CG iterations, Adaptive is the number of sparse grid adaptation iterations, and CP is the number of collocation points.

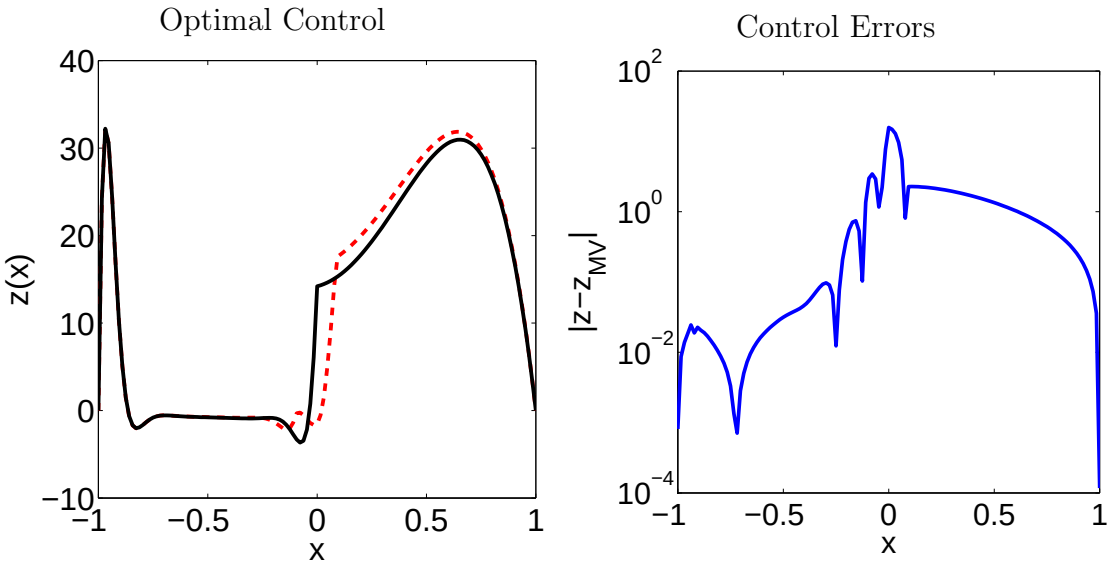
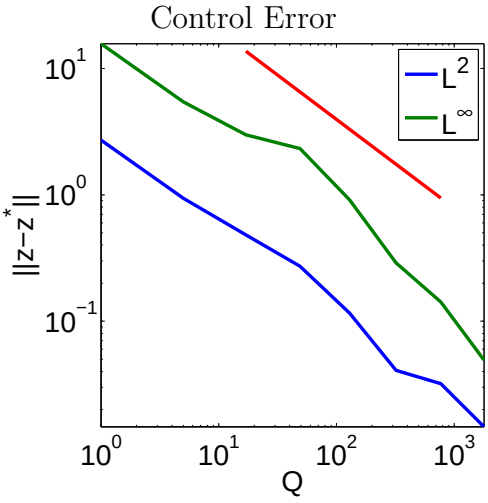


Figure 7.8: (Left) The optimal controls for the deterministic problem with $y \in \Gamma$ replaced by $\bar{y} = E[y]$ (solid black line). The red dashed line is the control computed via the stochastic problem. (Right) Errors between the optimal controls for the stochastic problem and the optimal controls for the mean value problem.



ℓ	Q	Error	Rate
0	1	2.71021418e+00	
1	5	9.39152597e-01	0.659
2	17	4.81411626e-01	0.546
3	49	2.72342525e-01	0.538
4	129	1.16051199e-01	0.881
5	321	4.08098085e-02	1.146
6	769	3.20679403e-02	0.276
7	1793	1.46056223e-02	0.929

Figure 7.9: (Left) Collocation error in the optimal controls. The least squares fitted convergence rate is $\nu = 0.7$. (Right) L^2 error and associated rate of decrease for the optimal controls.

7.2 Source Inversion Under Uncertainty

Source inversion problems are of paramount importance to many fields such as the monitoring of carbon output. In the deterministic setting, the source inversion problem is to determine the location and magnitude of the sources from observed data. Typically, these observations are point observations and the physics governing the system are assumed to be known. I will consider a steady state source inversion problem where the governing physics are advection-diffusion. Let $D = (0,1)^d \subset \mathbb{R}^d$ for $d = 2, 3$ be the physical domain. The PDE describing the physical system is

$$\begin{aligned} -\epsilon\Delta v + \mathbf{V}\cdot\nabla v &= z \quad \text{in } D \\ \nabla v \cdot n &= 0 \quad \text{on } \partial D_{\mathrm{N}} \\ v &= g \quad \text{on } \partial D_{\mathrm{D}} \end{aligned} \tag{7.2.1}$$

where ∂D_N denotes the outflow boundary and ∂D_D denotes the Dirichlet boundary. The source inversion problem with Tikhonov regularization can be written as

$$\min_{z\in\mathcal{Z}} \widehat{j}(z) := \frac{1}{2}\sum_{k=1}^{\mathcal{O}} |v(x_k;z) - \bar{v}_k|^2 + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2. \tag{7.2.2}$$

where $v = v(z)$ solves (7.2.1). This optimization problem is an instance of the test problem from Section 2.1 with $\mathcal{Z} = L^2(D)$, $\mathcal{V} = H^\ell(D)$, $\mathcal{W} = \mathcal{V}^*$, $\mathcal{H} = \mathbb{R}^{\mathcal{O}}$ with the Euclidean inner product, $\bar{q} = \bar{v} \in \mathbb{R}^{\mathcal{O}}$, and $\mathbf{Q} \in \mathcal{L}(H^\ell(D), \mathbb{R}^N)$ is defined as the point evaluation operator

$$\mathbf{Q}v = (v(x_1), \ldots, v(x_{\mathcal{O}}))^\top.$$

Note that $\mathbf{Q}$ is a continuous linear operator if $\ell > 0$ and $-\ell < -d + \frac{d}{2}$. $\mathbf{Q}$ is also continuous if $\mathcal{V} = H^1(D) \cap C^0(D)$. I will pose the optimization problem in $\mathcal{V} = H^1(D)$ and discretize (7.2.1) using continuous finite elements. Therefore, the discretized solution of (7.2.1) is a member of $H^1(D) \cap C^0(D)$ and $\mathbf{Q}$ is well defined.

A popular alternative to solving the deterministic problem (7.2.2), is to use Bayesian inference. In the Bayesian framework, one assumes there is some statistical

noise associated with the model (7.2.1) and the observations, $\bar{v}$. Incorporating this noise and prior knowledge of the sources allows one to write down Bayes' rule for the conditional probability. This framework is highly subjective as the optimal solution is based on prior knowledge. To circumvent this subjectivity, I have formulated (7.2.2) as a statistical inverse problem by assuming random field advection coefficients, $\mathbf{V}$. This is a frequentist approach to statistical inverse problems for which I can apply my adaptive stochastic collocation and trust region framework. Furthermore, I will focus on the expected value risk measure, $\sigma(Y) = E[Y]$.

I discretize the advection-diffusion equation (7.2.1) using streamline upwind Petrov-Galerkin (SUPG) stabilized continuous piecewise linear finite elements built on a uniform mesh of quadrilaterals. My SUPG implementation is based on [47]. The side length of the quadrilaterals used will be denoted by h. Let $\mathcal{V}_h = \text{span}\{\phi_1, \ldots, \phi_N\} \subset \mathcal{V}$ denote the FEM basis. The linear operators $\mathbf{A}(y)$ and $\mathbf{B}(y)$ from Section 2.1 are $\mathbf{A}(y) \in \mathbb{R}^{N\times N}$ and $\mathbf{B} \in \mathbb{R}^{N\times N}$ and are defined via the SUPG finite element discretization as

$$\begin{aligned}\mathbf{A}_{ij}(y) &= \int_D \epsilon \nabla\phi_i(x)\cdot\nabla\phi_j(x) + \mathbf{V}(y,x)\cdot\nabla\phi_i(x)\phi_j(x)\mathrm{d}x \\ &\quad + \tau(h)\int_D (\mathbf{V}(y)\cdot\nabla\phi_i(x))(\mathbf{V}(y)\cdot\nabla\phi_j(x))\mathrm{d}x \\ \mathbf{B}_{ij}(y) &= -\int_D \phi_i(x)\phi_j(x)\mathrm{d}x - \tau(h)\int_D \phi_i(x)(\mathbf{V}(y)\cdot\nabla\phi_j(x))\mathrm{d}x.\end{aligned}$$

Here, $\tau = \tau(h)$ denotes the SUPG parameter. Furthermore, the Riesz operator $\mathbf{R}$ corresponding to the inner product on $\mathcal{Z}$ is written as $\mathbf{R} \in \mathbb{R}^{N\times N}$ such that

$$\mathbf{R}_{ij} = \int_D \phi_i(x)\phi_j(x)\mathrm{d}x.$$

The inverse problem (7.2.2) is typically mesh dependent and convergence varies according to mesh size. To overcome this mesh dependence, I use the discretized objective function

$$\widehat{J}_{Qh}(z) = \frac{h^2}{2}\sum_{k=1}^{\mathcal{O}} E_Q[|u_h(x_k;z) - \bar{v}_k|^2] + \frac{\alpha}{2}\|z\|_{\mathcal{Z}}^2$$

where E_Q denotes the quadrature approximation to the expected value, $z = \sum_{n=1}^{N} z_n \phi_n$ for $\vec{z} = (z_1, \ldots, z_N)^\top \in \mathbb{R}^N$, and $u_h(y) = u_h(y; z) = \sum_{n=1}^{N} u_n \phi_n$ and $\vec{u}_h = (u_1, \ldots, u_N)^\top \in \mathbb{R}^N$ solves

$$\mathbf{A}(y)\vec{u}_h(y) + \mathbf{B}(y)\vec{z} = 0.$$

The scaling h^2 to the mismatch term in the objective function gives a scale independent regularization term.

I have implemented this problem using the Euclidean inner product and the $L^2(D)$ inner product defined on the discretized control space. The $L^2(D)$ inner product on the discretized control space corresponds to the inner product

$$\langle z, s \rangle_{\mathcal{Z}} = \vec{z}^\top \mathbf{R} \vec{s}$$

where $z = \sum_{n=1}^{N} z_n \phi_n$ and $s = \sum_{n=1}^{N} s_n \phi_n$ for $\vec{z} = (z_1, \ldots, z_N)^\top$ and $\vec{s} = (s_1, \ldots, s_N)^\top \in \mathbb{R}^N$. In the $L^2(D)$ inner product the gradient is

$$\nabla \widehat{J}_{Qh}(z) = \alpha \vec{z} + \mathbf{R}^{-1} E_Q[\mathbf{B}^* \vec{p}_h],$$

while in the Euclidean inner product the gradient is

$$\nabla \widehat{J}_{Qh}(z) = \alpha \mathbf{R} \vec{z} + E_Q[\mathbf{B}^* \vec{p}_h].$$

In both cases, $\vec{p}_h$ denotes the vector of coefficients corresponding to the solution to the finite element approximation to the adjoint equation

$$\mathbf{A}^*(y)\vec{p}_h + (\mathbf{Q}^*\mathbf{Q}\vec{u}_h - \bar{v}) = 0$$

where $\mathbf{Q} \in \mathbb{R}^{\mathcal{O} \times N}$ is the observation operator (i.e. part of an identity matrix scaled by h if the observations $\{x_k\}_{k=1}^{\mathcal{O}}$ correspond to mesh vertices).

7.2.1 Two Dimensional Source Inversion

In two dimensions ($d = 2$), the Neumann and Dirichlet boundaries are

$$\partial D_N = \{1\} \times [0, 1] \quad \text{and} \quad \partial D_D = \partial D \setminus \partial D_N.$$

and the inhomogeneous Dirichlet conditions are

$$g(y,x) = 0 \text{ for } x \in (0,1) \times \{0,1\},$$

and on $\{x \in \partial D_D \ : \ x \in \{0\} \times (0,1)\}$

$$g(y,x) = d_1(y_1, x_2)$$

where $d_1 : [-1,1] \times (0,1) \to [0,1]$ is defined as

$$d_1(\gamma,\zeta) = \begin{cases} 0 & \text{if } \zeta \notin (0.25, 0.75), \\ \sin(2\pi(\zeta - 0.25)) & \text{if } \zeta \in (0.25, 0.75) \text{ and } \gamma = 0, \\ \sin\left(\pi \frac{\exp(4\gamma(\zeta-0.25))-1)}{\exp(2\gamma)-1}\right) & \text{if } \zeta \in (0.25, 0.75) \text{ and } \gamma \neq 0. \end{cases} \tag{7.2.3}$$

The diffusivity parameter is deterministic and set to $\epsilon \equiv 10^{-2}$ and the two dimensional random advection field is

$$\mathbf{V}(y,x) = \left(e^{-\frac{(x_2-\bar{x})^2}{\gamma_1^2}} + \sum_{k=3}^{M} \frac{\gamma_0 y_k}{k} e^{-\frac{(x_2-\bar{x})^2}{\gamma_k^2}}\right) \begin{bmatrix} \cos(\theta y_2) \\ \sin(\theta y_2) \end{bmatrix}$$

where $\gamma_0 = 0.05$, $\gamma_1 = 0.0833$, $\gamma_k = \frac{\gamma_1}{(k-1)^2}$ for $k = 3, \ldots, M$, $\bar{x} = 0.5$, and $\theta = \frac{\pi}{32}$. Furthermore, the random vector, $y \in \Gamma := [-1,1]^M$, is uniformly distributed with joint density, $\rho(y) \equiv \frac{1}{2^M}$.

The true sources and observed data are depicted in Figure 7.10. For this example, there are fifteen true sources that are randomly distributed throughout the subdomain $[0.3,1] \times [0,1]$ with random magnitudes and widths. The observed state is computed by solving the state equation with the random variable $y \in \Gamma$ replaced with $y = E[y] = 0$. Point observations of the observed state are taken at each mesh vertex. The computational mesh used is a uniform 64 by 64 mesh of quadrilaterals. The observed state and stochastic state are solved on this mesh using continuous piecewise linear finite elements. For the numerical results presented here, $M = 5$ and the collocation points are taken to be level four isotropic Smolyak sparse grid knots built on one dimensional Clenshaw-Curtis interpolation knots. Figure 7.11 depicts the

computed sources and the resulting expected value of the state. Both Figure 7.10 and Figure 7.11 only display the sources in the subdomain $[0.3, 1] \times [0, 1]$. The computed sources in the full domain are depicted in Figure 7.12. Notice that there are oscillations near the Dirichlet boundary $x_1 = 0$. This phenomenon is due the discrepancy between the inhomogeneous Dirichlet conditions $E[g(y, x)]$ and $g(0, x)$ (i.e. the discrepancy between the stochastic state equation and the deterministic observed state equation). The difference in the Dirichlet boundary conditions can be seen in the top right image in Figure 7.12. One will also notice in the bottom image of Figure 7.12 that the uncertain Dirichlet conditions are a main source of variability in the state equation. The adapted sparse grid index set for this 2D source inversion problem is displayed in Figure 7.13. Figure 7.13 characterizes the anisotropy associated with the random variables in this problem. The top left image shows isotropy and a strong dependence on the first two directions y_1 and y_2. Directions y_1 and y_2 obtain the maximum level for the sparse grid (i.e. $\ell = 5$). These directions correspond to the Dirichlet condition and the angle of advection. The subsequent images demonstrate the decreasing importance of the random variables y_3, y_4, and y_5, which all correspond to the advection amplitude.

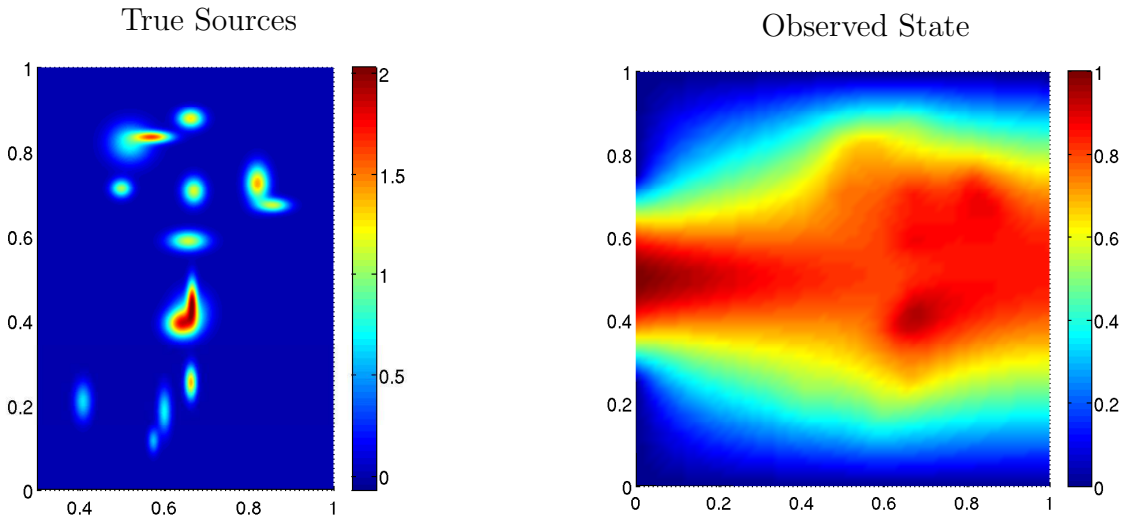


Figure 7.10: (Left) True sources. (Right) Observed state computed by solving the state equation with $y = 0$.

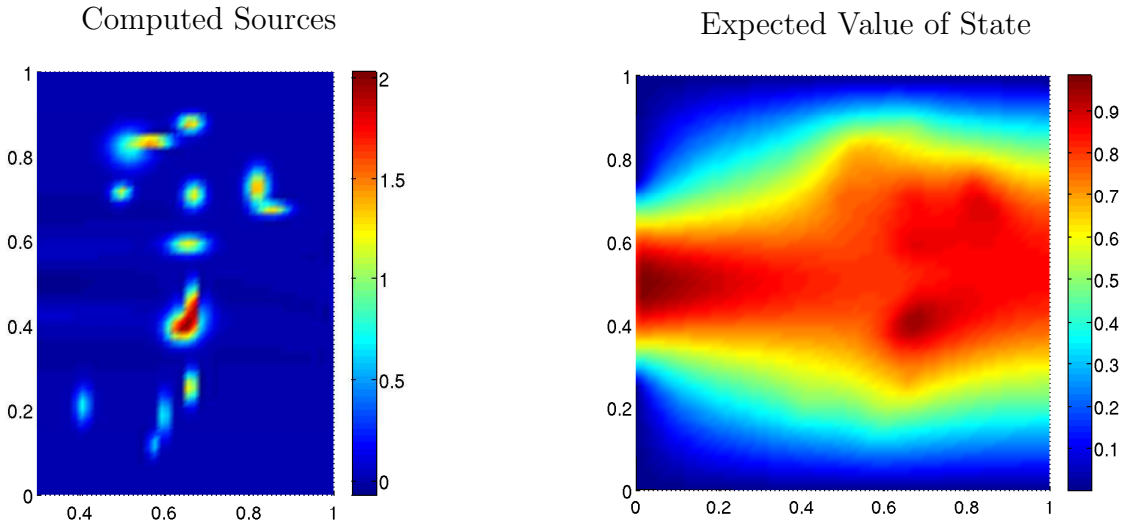


Figure 7.11: (Left) Computed sources. (Right) Expected value of optimal state.

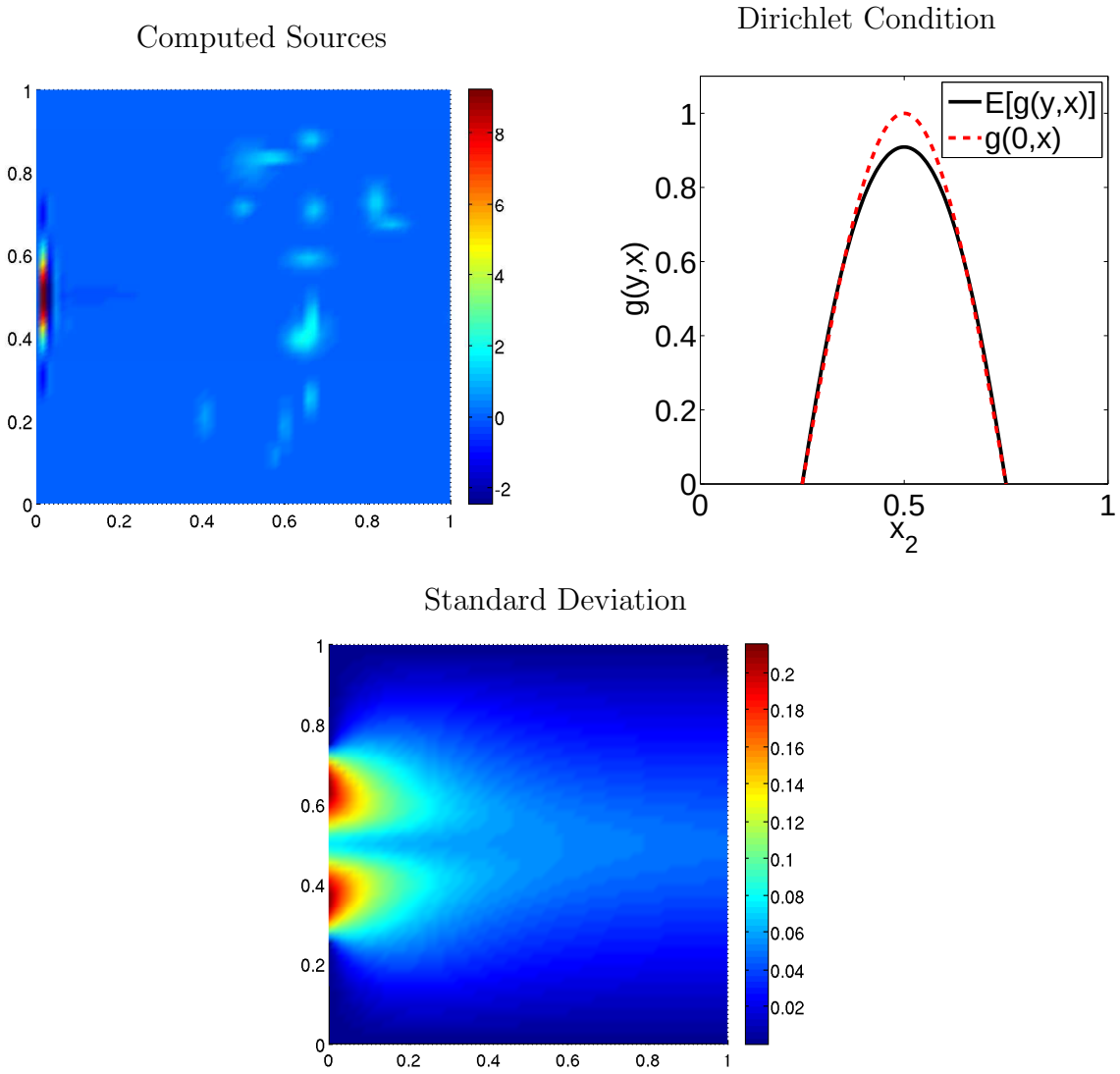


Figure 7.12: (Left) Computed sources. (Right) The expected value of the inhomogeneous Dirichlet condition (black) and the Dirichlet condition evaluated at $y = 0$. This difference accounts for the hot spot near the boundary $x_1 = 0$ in the left image. (Bottom) Standard deviation of the state. Notice that most variation is due to Dirichlet conditions.

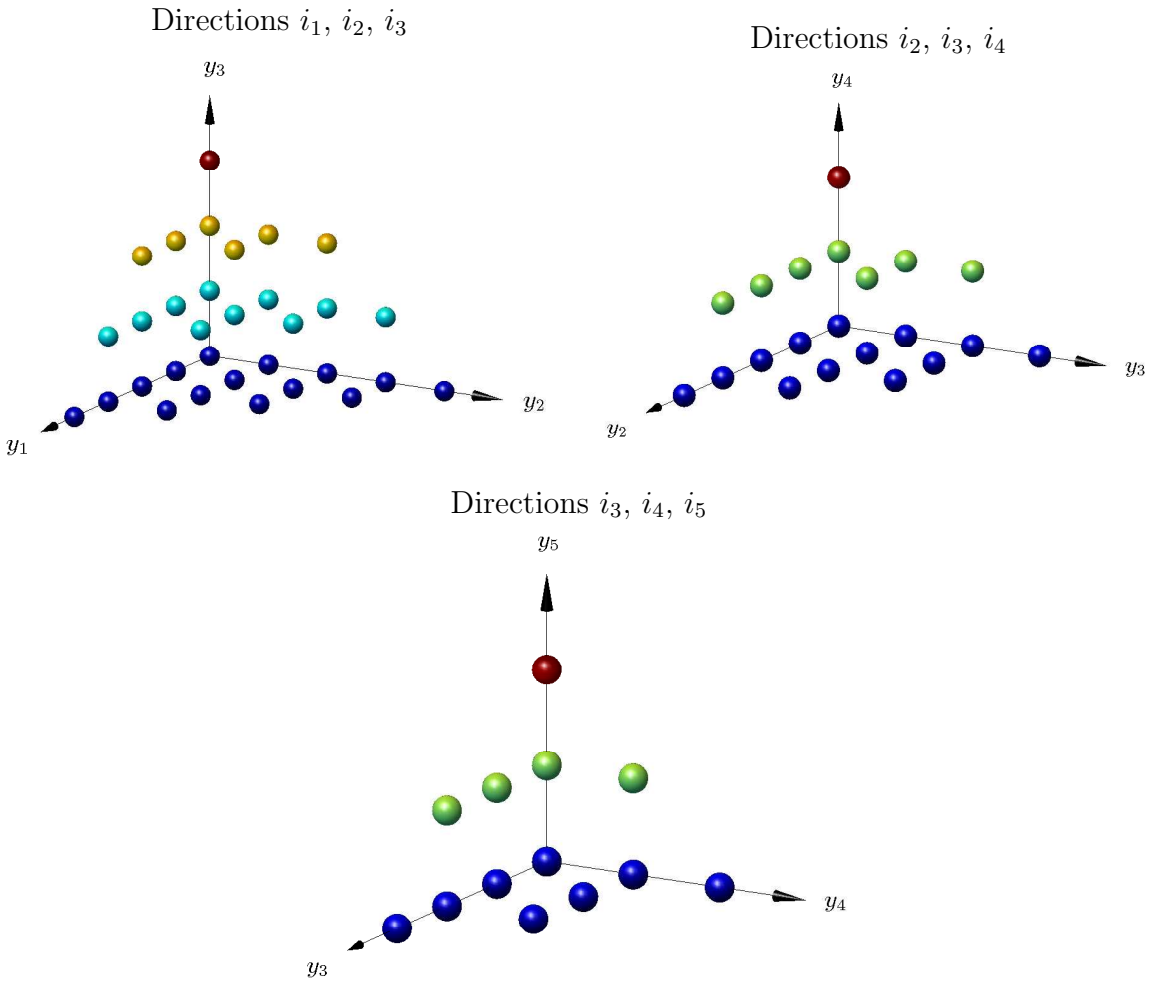


Figure 7.13: The final adapted sparse grid index set for 2D source inversion.

7.2.2 Three Dimensional Source Inversion

In three dimensions ($d = 3$), the Neumann and Dirichlet boundaries are

$$\partial D_N = \{1\} \times [0,1] \times [0,1] \quad \text{and} \quad \partial D_D = \partial D \setminus \partial D_N.$$

and the inhomogeneous Dirichlet conditions are

$$g(y,x) = 0 \text{ for } x \in (0,1) \times (0,1) \times \{0,1\} \cup (0,1) \times \{0,1\} \times (0,1),$$

and on $\{x \in \partial D_D \ : \ x \in \{0\} \times (0,1) \times (0,1)\}$

$$g(y,x) = d_1(y_1, x_2) d_1(y_2, x_3)$$

where $d_1(\gamma, \zeta)$ is defined in (7.2.3). The diffusivity parameter is deterministic and set to $\epsilon \equiv 10^{-2}$ and the two dimensional random advection field is

$$\mathbf{V}(y,x) = \left(e^{-\frac{(x_2 - \bar{x})^2}{\gamma_1^2}} + \sum_{k=4}^{M} \frac{\gamma_0 y_k}{k} e^{-\frac{(x_2 - \bar{x})^2}{\gamma_k^2}} \right) \begin{bmatrix} \cos(\theta y_3) \\ \sin(\theta y_3) \\ 0 \end{bmatrix}$$

where $\gamma_0 = 0.05$, $\gamma_1 = 0.0833$, $\gamma_k = \frac{\gamma_1}{(k-2)^2}$ for $k = 4, \ldots, M$, $\bar{x} = 0.5$, and $\theta = \frac{\pi}{32}$. Furthermore, the random vector, $y \in \Gamma := [-1,1]^M$, is uniformly distributed with joint density, $\rho(y) \equiv \frac{1}{2^M}$.

The true sources and observed data are depicted in Figure 7.14. As in the 2D source inversion example, there are fifteen true sources that are randomly distributed throughout the subdomain $[0.3,1] \times [0,1] \times [0,1]$ with random magnitudes and widths. The observed state is computed by solving the state equation with the random variable $y \in \Gamma$ replaced with $y = E[y] = 0$. Point observations of the observed state are taken at each mesh vertex. The computational mesh used is a uniform 32 by 32 by 32 mesh of hexahedron. The observed state and stochastic state are solved on this mesh using continuous piecewise linear finite elements. For the numerical results presented here, $M = 6$ and the collocation points are taken to be level two isotropic Smolyak sparse grid knots built on one dimensional Clenshaw-Curtis interpolation knots. Figure 7.15

depicts the computed sources in the upper right image, contour lines of the expected value of the state, and contour lines of the standard deviation of the state. Notice that there are oscillations near the Dirichlet boundary $x_1 = 0$ as was the case in the 2D source inversion example. This phenomenon is due the discrepancy between the inhomogeneous Dirichlet conditions $E[g(y, x)]$ and $g(0, x)$ (i.e. the discrepancy between the stochastic state equation and the deterministic observed state equation).

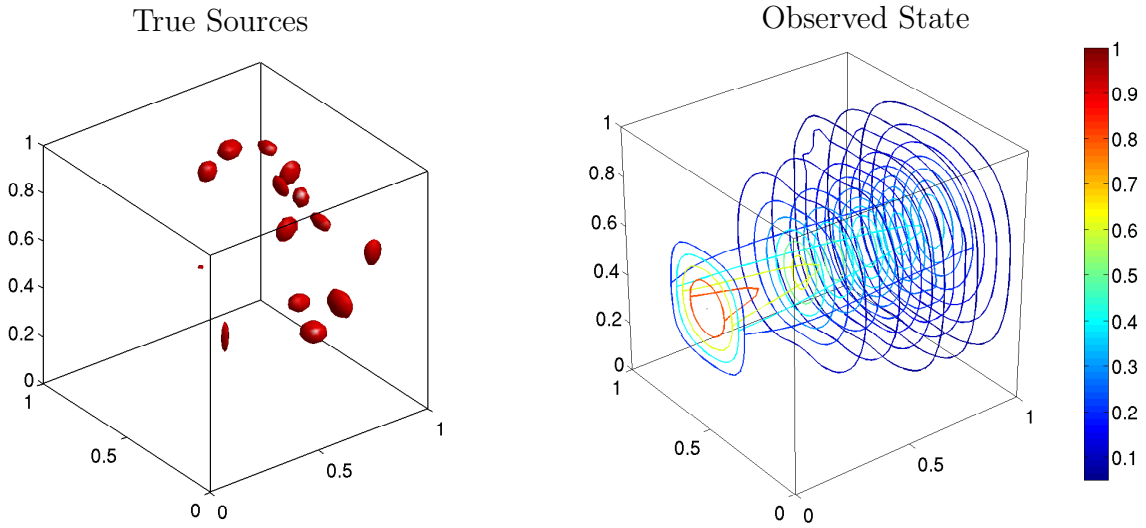


Figure 7.14: (Left) True sources. These sources are plotted using an isosurface of $z = 0.2$. (Right) Observed state computed by solving the state equation with $y = 0$.

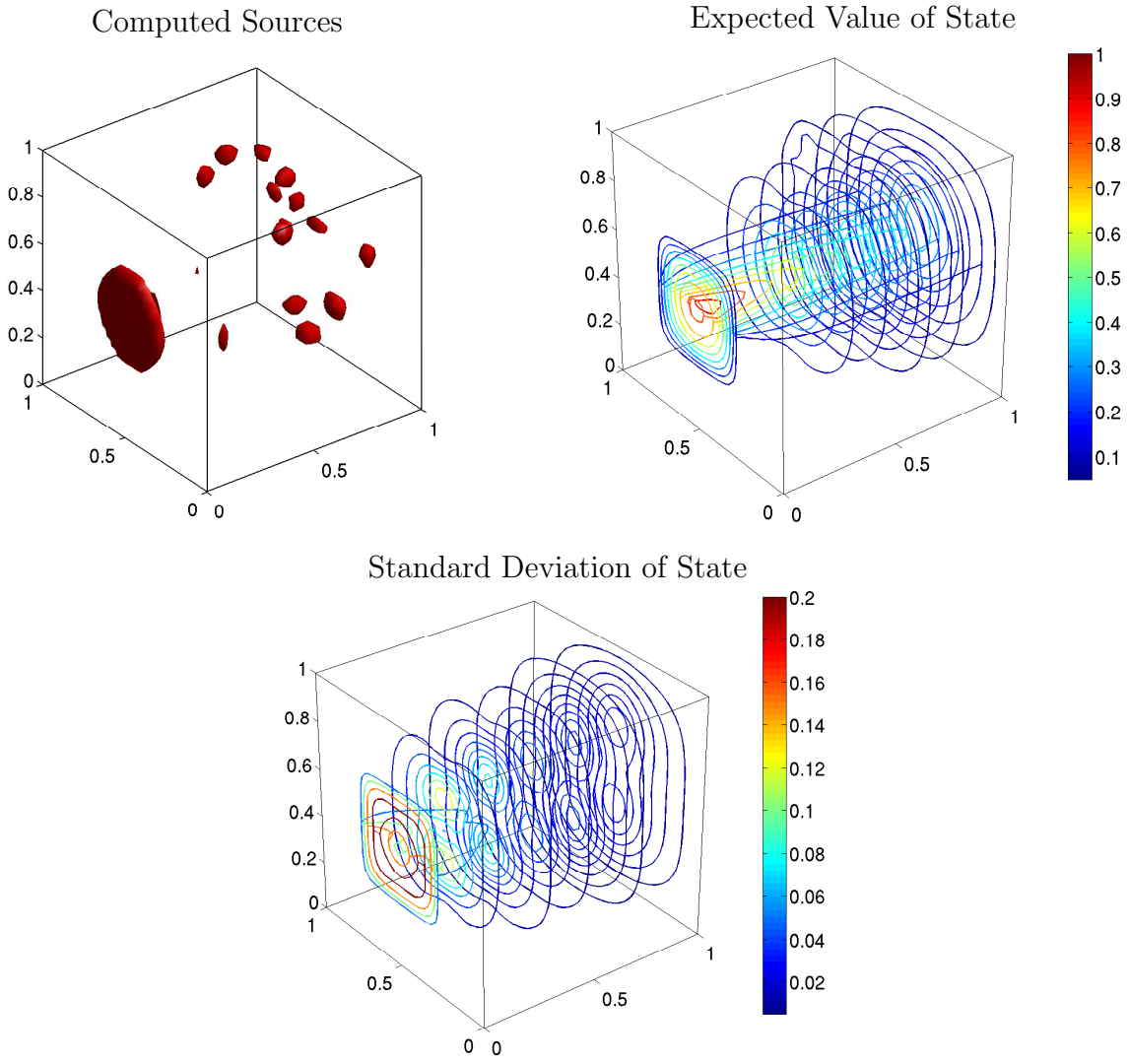


Figure 7.15: Isosurface of $z = 0.2$ for the computed sources (left), contours of the expected value of the optimal state (right), and contours of the standard deviation of the optimal state (bottom). Notice the phenomenon near the inhomogeneous Dirichlet boundary in the upper left figure. This "fake source" is due to the difference between $E[g(y,x)]$ and $g(0,x)$ as in the 2D source inversion example.

Chapter 8

Conclusions and Future Work

I have presented in the thesis a general problem formulation for equality constrained optimization problems where the equality constraint depends on random inputs. This general formulation includes many interesting optimization problems such as the risk-averse optimal control and design of PDEs with uncertain coefficients. The main contributions of my work are the analysis and algorithms developed for the risk-averse optimization of PDEs with uncertain coefficients. I have developed a sparse grid stochastic collocation discretization scheme for these optimization problems and extended error bounds from the theory of stochastic collocation for PDEs with uncertain coefficients to the case of optimization. Furthermore, I have employed generalized sparse grids to obtain efficient discretizations of these optimization problems and have proven certain interpolation and approximation properties for these generalized sparse grids operators. I have developed an adaptive stochastic collocation framework for the efficient solution of optimization problems governed by PDEs with uncertain coefficients. This adaptive framework utilizes adaptive sparse grids to generate inexpensive approximate models and guides adaptivity using the trust region algorithm. I have applied this framework using both the basic trust region algorithm and the retrospective trust region algorithm. For the retrospective trust region algorithm, I have proven global first order convergence under a weakened condition on gradient

exactness. Finally, I have implemented this trust region framework employing truncated conjugate gradients (CG) to solve the trust region subproblem. To increase the efficiency of CG and further reduce the number of PDE solves required at every trust region iteration, I employ automatic preconditioning using limited memory BFGS inverse Hessian approximations.

This thesis is focused on the efficient solution of the reduced space problem (2.2.1) using adaptive sparse grid collocation. The trust region algorithm that I have developed here is not limited to adaptivity in the stochastic dimension, but can also be extended to spatial adaptivity via finite elements and even model order reduction adaptivity for time dependent problems via adaptive snapshot selection for projection based reduced order modeling. Although the method I have developed is tied to non-intrusive methods, one may be able to extend adaptive polynomial chaos and Taylor series approximation methods to the optimization context using this trust region framework. These additional outlets for adaptivity are yet to be studied and are natural extensions of my doctoral work.

The incorporation of risk measures brings about many issues for analysis, computation, and algorithms, but allows for explicit handling of the risk or variation associated with each design or control. Risk measures are a natural means of quantifying tail value risk in engineering design and safety analysis where the goal is to determine a design that will withstand extreme and rare events. Convergence analysis is a first step to incorporating risk measures in an optimization scheme. It is essential to know that the controls computed via a discretization of (2.2.1) do converge to the true controls as the discretizations are refined. This is a challenging task as most risk measures are not Fréchet differentiable. This convergence analysis can be performed by substituting a smooth approximation of the risk measure. Tracking these discretization and approximation errors through to the optimal controls is essential for accurate quantification of risk and uncertainty.

In addition to risk measures, I would like to incorporate general constraints into

my optimization formulation. In particular, I would like to tackle chance constraints. Chance constraints offer a natural means of accurately estimating operating ranges and fault tolerances for certifying engineering components in control and design problems. As mentioned earlier, chance constraints may cause issues for gradient based algorithms since they are not necessarily Fréchet differentiable. These constraints can be handled in a similar way to risk measure via smooth approximation. Again, these approximation errors must be tracked through to the optimal control values. The addition of general constraints is beyond the scope of my trust region framework. The algorithm I have proposed works for unconstrained reduced space problems, (2.2.1). In the presence of state and control constraints, my trust region algorithm must be extended. I plan to extend my algorithmic capabilities to the full space and determine an appropriate manner of incorporating adaptivity. Such adaptive full space algorithms have been considered in the case of finite element adaptivity in [127].

Risk measures and chance constraints are problem formulation issues and are dictated by the desired application. A main concern of these optimization problems is algorithmic efficiency. In the case of time dependent problems, some form of model order reduction is critical in making the numerical solution of such problems feasible. I am interested in coupling projection based model order reduction techniques with my stochastic collocation framework. In this case, a fixed projection basis (i.e. fixed snapshots) may not be necessary. In fact, the basis can be built adaptively in the same manner as the sparse grid using the inexact gradient condition, (5.1.3). Another approach to incorporating model order reduction is to fix the projection basis by a greedy sampling of the parameter space to choose "optimal" snapshots [31].

Bibliography

[1] N. Agarwal and N. R. Aluru. A domain adaptive stochastic collocation approach for analysis of MEMS under uncertainties. *J. Comput. Phys.*, 228(20):7662–7688, 2009.

[2] N. Alexandrov and J. E. Dennis. Multilevel algorithms for nonlinear optimization. In J. Borggaard, J. Burkardt, M. D. Gunzburger, and J. Peterson, editors, *Optimal Design and Control*, pages 1–22, Basel, Boston, Berlin, 1995. Birkhäuser Verlag.

[3] N. Alexandrov, J. E. Dennis Jr., R. M. Lewis, and V. Torczon. A trust region framework for managing the use of approximation models in optimization. *Structural Optimization*, 15:16–23, 1998. Appeared also as ICASE report 97–50.

[4] H. Antil, M. Heinkenschloss, R. H. W. Hoppe, and D. C. Sorensen. Domain decomposition and model reduction for the numerical solution of PDE constrained optimization problems with localized optimization variables. *Computing and Visualization in Science*, 13:249–264, 2010. 10.1007/s00791-010-0142-4.

[5] Ph. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, 1999.

[6] V. I. Averbukh and O. G. Smolyanov. The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys*, 22(6):201–258, 1967.

[7] V. I. Averbukh and O. G. Smolyanov. The various definitions of the derivative in linear topological spaces. *Russian Mathematical Surveys*, 23(4):67–113, 1968.

[8] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Numer. Anal.*, 45(3):1005–1034 (electronic), 2007.

[9] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Rev.*, 52(2):317–355, 2010.

[10] I. Babuška, R. Tempone, and G. E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J. Numer. Anal.*, 42(2):800–825 (electronic), 2004.

[11] I. Babuška, R. Tempone, and G. E. Zouraris. Solving elliptic boundary value problems with uncertain coefficients by the finite element method: the stochastic formulation. *Comput. Methods Appl. Mech. Engrg.*, 194(12-16):1251–1294, 2005.

[12] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. Stochastic Galerkin and collocation methods for PDEs with random coefficients: a numerical comparison. 09-33, Institute for Computational Engineering and Science, University of Texas, Austin, Austin Texas, 2009. appeared as [14].

[13] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods. Technical Report MOX Report 23/2011, MOX Modeling and Scientific Computing, Politecnico di Milano, Italy, 2011.

[14] J. Bäck, F. Nobile, L. Tamellini, and R. Tempone. Stochastic spectral Galerkin and collocation methods for PDEs with random coefficients: A numerical comparison. In J. S. Hesthaven and E. M. Ronquist, editors, *Spectral and High Order*

Methods for Partial Differential Equations, Lecture Notes in Computational Science and Engineering, Vol. 76, pages 43–62. Springer Berlin Heidelberg, 2011. 10.1007/978-3-642-15337-2_3.

[15] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000. Multivariate polynomial interpolation.

[16] F. Bastin, C. Cirillo, and Ph. L. Toint. An adaptive Monte Carlo algorithm for computing mixed logit estimators. *Comput. Manag. Sci.*, 3(1):55–79, 2006.

[17] F. Bastin, V. Malmedy, M. Mouffe, Ph. L. Toint, and D. Tomanos. A retrospective trust-region method for unconstrained optimization. *Math. Program.*, 123(2, Ser. A):395–418, 2010.

[18] R. Becker, M. Braack, D. Meidner, R. Rannacher, and B. Vexler. Adaptive finite element methods for PDE-constrained optimal control problems. In W. Jäger, R. Rannacher, and J. Warnatz, editors, *Reactive flows, diffusion and transport*, pages 177–205. Springer, Berlin, 2007.

[19] R. Becker, H. Kapp, and R. Rannacher. Adaptive finite element methods for optimal control of partial differential equations: Basic concepts. *SIAM J. Control and Optimization*, 39:113–132, 2000.

[20] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.*, 10:1–102, 2001.

[21] P. Benner, V. Mehrmann, and D. C. Sorensen, editors. *Dimension Reduction of Large-Scale Systems*. Lecture Notes in Computational Science and Engineering, Vol. 45. Springer-Verlag, Heidelberg, 2005.

[22] L. Biegler, G. Biros, O. Ghattas, M. Heinkenschloss, D. Keyes, B. Mallick, Y. Marzouk, L. Tenorio, B. van Bloemen Waanders, and K. Willcox, editors.

Large-Scale Inverse Problems and Quantification of Uncertainty. John Wiley & Sons, Ltd, Chichester, 2011.

[23] G. Biros and O. Ghattas. Parallel Lagrange–Newton–Krylov–Schur methods for PDE–constrained optimization. part I: The Krylov–Schur solver. *SIAM J. Sci. Comput.*, 27(2):587–713, 2005.

[24] A. Borzì. Multigrid and sparse-grid schemes for elliptic control problems with random coefficients. *Comput. Vis. Sci.*, 13:153–160, 2010.

[25] A. Borzì and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. Computational Science and Engineering, Vol. 8. SIAM, Philadelphia, 2012.

[26] A. Borzì, V. Schulz, C. Schillings, and G. von Winckel. On the treatment of distributed uncertainties in PDE constrained optimization. *GAMM Mitteilungen*, 33(2):230–246, 2010.

[27] A. Borzì and G. von Winckel. Multigrid methods and sparse-grid collocation techniques for parabolic optimal control problems with random coefficients. *SIAM J. Sci. Comput.*, 31(3):2172–2192, 2009.

[28] A. Borzì and G. von Winckel. A POD framework to determine robust controls in pde optimization. *Comput. Vis. Sci.*, 14:91–103, 2011.

[29] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, Berlin, Heidelberg, New York, second edition, 2002.

[30] L. Brutman. On the Lebesgue function for polynomial interpolation. *SIAM Journal on Numerical Analysis*, 15(4):pp. 694–704, 1978.

[31] T. Bui-Thanh, K. Willcox, O. Ghattas, and B. van Bloemen Waanders. Goal-oriented, model-constrained optimization for for reduction of large-scale systems. *Journal of Computational Physics*, 224(2):880–896, 2007.

[32] H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[33] C. Carstensen. Some remarks on the history and future of averaging techniques in a posteriori finite element error analysis. *ZAMM Z. Angew. Math. Mech.*, 84(1):3–21, 2004.

[34] C. Carstensen. A unifying theory of a posteriori finite element error control. *Numer. Math.*, 100(4):617–637, 2005.

[35] C. Carstensen, R. H. W. Hoppe, C. Löbhard, and M. Eigel. A review of unified a posteriori finite element error control. Technical Report 2338, Intitute for Mathematics and its Applications, University Minnesota, October 2010.

[36] R. G. Carter. On the global convergence of trust region algorithms using inexact gradient information. *SIAM J. Numer. Anal.*, 28:251–265, 1991.

[37] B. Cengiz. On the duals of Lebesgue-Bochner L^p spaces. *Proc. Amer. Math. Soc.*, 114(4):923–926, 1992.

[38] C. H. Chen and O. L. Mangasarian. Smoothing methods for convex inequalities and linear complementarity problems. *Math. Programming*, 71(1, Ser. A):51–69, 1995.

[39] A. Chkifa, A. Cohen, R. DeVore, and C. Schwab. Sparse adaptive taylor approximation algorithms for parametric and stochastic elliptic pdes. SAM Research Report 2011–44, Seminar für Angewandte Mathematik, ETH Zürich, 2011.

[40] C. W. Clenshaw and A. R. Curtis. A method for numerical integration on an automatic computer. *Numer. Math.*, 2:197–205, 1960.

[41] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust–Region Methods*. SIAM, Philadelphia, 2000.

[42] B. J. Debusschere, H. N. Najm, P. P. Pébay, O. M. Knio, R. G. Ghanem, and O. P. Le Maître. Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM J. Sci. Comput.*, 26(2):698–719 (electronic), 2004.

[43] J. E. Dennis and V. Torczon. Approximation model managemet for optimization. In *Proceedings from the AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Work-in-progress Paper AIAA–96–4099-CP*, pages 1044–1046, 1996.

[44] J. E. Dennis and V. Torczon. Managing approximation models in optimization. In N. Alexandrov and M. Y. Hussaini, editors, *Multidisciplinary Design Optimization. State of the Art*, pages 330–347, Philadelphia, 1997. SIAM.

[45] R. A. DeVore and G. G. Lorentz. *Constructive Approximation*. Springer Verlag, New York, Berlin, Heidelberg, London, Paris, 1993.

[46] M. S. Eldred and J. Burkardt. Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification. In *paper AIAA-2009-0976 in Proceedings of the 47th AIAA Aerospace Sciences Meeting, Orlando, FL, Jan. 5-8, 2009*, 2009.

[47] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics*. Oxford University Press, Oxford, 2005.

[48] M. Fahl and E.W. Sachs. Reduced order modelling approaches to PDE–constrained optimization based on proper orthogonal decompostion. In L. T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders, editors, *Large-Scale PDE-Constrained Optimization*, Lecture Notes in Computational Science and Engineering, Vol. 30, Heidelberg, 2003. Springer-Verlag.

[49] G. B. Folland. *Real analysis. Modern techniques and their applications*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, second edition, 1999.

[50] J. Garcke and M. Griebel. Classification with sparse grids using simplicial basis functions. *Intell. Data Anal.*, 6:483–502, December 2002.

[51] A. Genz and B. D. Keister. Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight. *J. Comput. Appl. Math.*, 71(2):299–309, 1996.

[52] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3-4):209–232, 1998.

[53] T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71(1):65–87, 2003.

[54] R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach*. Springer-Verlag, New York, 1991.

[55] M. Griebel and M. Holtz. Dimension-wise integration of high-dimensional functions with applications to finance. *J. Complexity*, 26(5):455–489, 2010.

[56] M. Griebel and S. Knapek. Optimized general sparse grid approximation spaces for operator equations. *Math. Comp.*, 78(268):2223–2257, 2009.

[57] S. Gugercin and A. C. Antoulas. A survey of model reduction by balanced truncation and some new results. *Internat. J. Control*, 77(8):748–766, 2004.

[58] M. Hegland. Adaptive sparse grids. *ANZIAM J.*, 44((C)):C335–C353, 2002.

[59] M. Heinkenschloss. Numerical solution of implicitly constrained optimization problems. Technical Report TR08–05, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005–1892, 2008.

[60] M. Heinkenschloss and L. N. Vicente. Analysis of inexact trust–region SQP algorithms. *SIAM J. Optimization*, 12:283–302, 2001.

[61] M. Hintermüller and R. H. W. Hoppe. Goal-oriented adaptivity in control constrained optimal control of partial differential equations. *SIAM J. Control Optim.*, 47(4):1721–1743, 2008.

[62] M. Hintermüller, R. H. W. Hoppe, Y. Iliash, and M. Kieweg. An a posteriori error analysis of adaptive finite element methods for distributed elliptic control problems with control constraints. *ESAIM Control Optim. Calc. Var.*, 14(3):540–560, 2008.

[63] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with Partial Differential Equations*, volume 23 of *Mathematical Modelling, Theory and Applications*. Springer Verlag, Heidelberg, New York, Berlin, 2009.

[64] K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 1947(37):79, 1947.

[65] G. E. Karniadakis, C.-H. Su, D. Xiu, D. Lucor, C. Schwab, and R. A. Todor. Generalized polynomial chaos solution for differential equations with random inputs. Technical Report 2005-01, Seminar for Applied Mathematics, ETH Zurich, Zurich, Switzerland, 2005.

[66] L. Kaufman. Reduced storage, quasi-Newton trust region approaches to function optimization. *SIAM J. Optim.*, 10(1):56–69 (electronic), 1999.

[67] D. P. Kouri. Optimization governed by stochastic partial differential equations. Master's thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2010. Available as CAAM TR10-20.

[68] A. Kunoth and Ch. Schwab. Analytic regularity and GPC approximation for control problems constrained by linear parametric elliptic and parabolic PDEs.

SAM Research Report 2011–54, Seminar für Angewandte Mathematik, ETH Zürich, 2011.

[69] P. D. Lax. *Functional Analysis*. John Wiley & Sons, New-York, Chicester, Brisbane, Toronto, 2002.

[70] O. P. Le Maitre and O. M. Knio. *Spectral Methods for Uncertainty Quantification With Applications to Computational Fluid Dynamics*. Scientific Computation. Springer-Verlag, Berlin, 2010.

[71] M. Loève. Fonctions aléatoires de second ordre. *Revue Sci.*, 84:195–206, 1946.

[72] D. Lucor and G. E. Karniadakis. Adaptive generalized polynomial chaos for nonlinear random oscillators. *SIAM J. Sci. Comput.*, 26(2):720–735 (electronic), 2004.

[73] X. Ma and N. Zabaras. An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations. *J. Comput. Phys.*, 228(8):3084–3113, 2009.

[74] X. Ma and N. Zabaras. An efficient Bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method. *Inverse Problems*, 25(3):035013, 27, 2009.

[75] K. Marti. *Stochastic optimization methods*. Springer-Verlag, Berlin, 2005.

[76] Y. M. Marzouk and D. Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. *Commun. Comput. Phys.*, 6(4):826–847, 2009.

[77] D. Meidner and B. Vexler. Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.*, 46(1):116–142 (electronic), 2007.

[78] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems. I. Problems without control constraints. *SIAM J. Control Optim.*, 47(3):1150–1177, 2008.

[79] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems. II. Problems with control constraints. *SIAM J. Control Optim.*, 47(3):1301–1329, 2008.

[80] J. L. Morales and J. Nocedal. Automatic preconditioning by limited memory quasi-Newton updating. *SIAM J. Optim.*, 10(4):1079–1096 (electronic), 2000.

[81] J. J. Moré. Recent developments in algorithms and software for trust region methods. In A. Bachem, M. Grötschel, and B. Korte, editors, *Mathematical Programming, The State of The Art*, pages 258–287. Springer Verlag, Berlin, Heidelberg, New-York, 1983.

[82] J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM J. Sci. Statist. Comput.*, 4(3):553–572, 1983.

[83] H. N. Najm. Uncertainty quantification and polynomial chaos techniques i in computational fluid dynamics. *Annual Review of Fluid Mechanics*, 41:35–52, 2009.

[84] H. N. Najm, B. J. Debusschere, Y. M. Marzouk, S. Widmer, and O. P. Le Maitre. Uncertainty quantification in chemical systems. *International Journal for Numerical Methods in Engineering*, 80(6-7):789–814, 2009.

[85] F. Nobile, R. Tempone, and C. G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2411–2442, 2008.

[86] F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.

[87] E. Novak and K. Ritter. High-dimensional integration of smooth functions over cubes. *Numer. Math.*, 75(1):79–97, 1996.

[88] E. Novak and K. Ritter. Simple cubature formulas with high polynomial exactness. *Constr. Approx.*, 15(4):499–522, 1999.

[89] E. Novak and H. Woźniakowski. *Tractability of multivariate problems. Volume II: Standard information for functionals*, volume 12 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010.

[90] B.K. Pagnoncelli, S. Ahmed, and A. Shapiro. Sample average approximation method for chance constrained programming: theory and applications. *J. Optim. Theory Appl.*, 142(2):399–416, 2009.

[91] T. N. L. Patterson. The optimum addition of points to quadrature formulae. *Math. Comp. 22 (1968), 847–856; addendum, ibid.*, 22(104, loose microfiche supp.):C1–C11, 1968.

[92] K. Petras. On the Smolyak cubature error for analytic functions. *Adv. Comput. Math.*, 12(1):71–93, 2000. High dimensional integration.

[93] K. Petras. Smolyak cubature of given polynomial degree with few nodes for increasing dimension. *Numer. Math.*, 93(4):729–753, 2003.

[94] M. J. D. Powell. Convergence properties of a class of minimization algorithms. In O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, editors, *Nonlinear Programming 2*, pages 1–27, Boston, New-York, London,..., 1975. Academic Press.

[95] R. Rannacher and B. Vexler. A priori error estimates for the finite element discretization of elliptic parameter identification problems with pointwise measurements. *SIAM J. Control Optim.*, 44(5):1844–1863 (electronic), 2005.

[96] D. Ridzal. *Trust Region SQP Methods With Inexact Linear System Solves For Large-Scale Optimization.* PhD thesis, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2006. Available as CAAM TR06–02.

[97] T. J. Rivlin. The Lebesgue constants for polynomial interpolation. In H. Garnir, K. Unni, and J. Williamson, editors, *Functional Analysis and its Applications*, volume 399 of *Lecture Notes in Mathematics*, pages 422–437. Springer Berlin / Heidelberg, 1974.

[98] R. T. Rockafellar. Coherent approaches to risk in optimization under uncertainty. *Tutorials in Operations Research INFORMS*, pages 38–61, 2007.

[99] A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Math. Oper. Res.*, 31(3):433–452, 2006. see also [101].

[100] A. Ruszczyński and A. Shapiro. Optimization of risk measures. In G. Calafiore and F. Dabbene, editors, *Probabilistic and Randomized Methods for Design Under Uncertainty*, pages 119–157, London, 2006. Springer Verlag.

[101] A. Ruszczyński and A. Shapiro. Corrigendum to: "Optimization of convex risk functions" [Math. Oper. Res. **31** (2006), no. 3, 433–452]. *Math. Oper. Res.*, 32(2):496, 2007.

[102] R. A. Ryan. *Introduction to tensor products of Banach spaces.* Springer Monographs in Mathematics. Springer-Verlag London Ltd., London, 2002.

[103] P. Sarma, L. Durlofsky, and K. Aziz. Kernel principal component analysis for efficient, differentiable parameterization of multipoint geostatistics. *Math. Geosci.*, 40(1):3–32, 2008.

[104] C. Schillings. *Optimal Aerodynamic Design under Uncertainties*. PhD thesis, Fb–IV, Mathematik, Universität Trier, D–54286 Trier, Germany, 2010.

[105] V. Schulz and C. Schillings. On the nature and treatment of uncertainties in aerodynamic design. *AIAA Journal*, 47(3):646–654, 2009.

[106] C. Schwab and R. A. Todor. Karhunen-Loève approximation of random fields by generalized fast multipole methods. *J. Comput. Phys.*, 217(1):100–122, 2006.

[107] A. Shapiro. On concepts of directional differentiability. *J. Optim. Theory Appl.*, 66(3):477–487, 1990.

[108] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, 2009.

[109] G. A. Shultz, R. B. Schnabel, and R. H. Byrd. A family of trust region based algorithms for unconstrained minimization with strong global convergence properties. *SIAM J. Numer. Anal.*, 22:47–67, 1985.

[110] S. A. Smoljak. Quadrature and interpolation formulae on tensor products of certain function classes. *Dokl. Akad. Nauk SSSR*, 148:1042–1045, 1963. Russian.

[111] S. A. Smoljak. Quadrature and interpolation formulae on tensor products of certain function classes. *Soviet Math. Dokl.*, 4:240–243, 1963.

[112] A. H. Stroud. *Approximate calculation of multiple integrals*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1971. Prentice-Hall Series in Automatic Computation.

[113] R. A. Todor and C. Schwab. Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA J. Numer. Anal.*, 27(2):232–261, 2007.

[114] S. Uryas′ev. Derivatives of probability functions and integrals over sets given by inequalities. *J. Comput. Appl. Math.*, 56(1-2):197–223, 1994. Stochastic programming: stability, numerical methods and applications (Gosen, 1992).

[115] S. Uryasev. Derivatives of probability functions and some applications. *Ann. Oper. Res.*, 56:287–311, 1995. Stochastic programming (Udine, 1992).

[116] S. Uryasev, editor. *Probabilistic constrained optimization. Methodology and applications*, volume 49 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 2000.

[117] S. Uryasev and R. T. Rockafellar. Conditional value-at-risk: optimization approach. In S. Uryasev and P. M. Pardalos, editors, *Stochastic optimization: algorithms and applications. Papers from the conference held at the University of Florida, Gainesville, FL, February 20–22, 2000*, volume 54 of *Appl. Optim.*, pages 411–435. Kluwer Acad. Publ., Dordrecht, 2001.

[118] C. R. Vogel. *Computational Methods for Inverse Problems*. Frontiers in Applied Mathematics, Vol 24. SIAM, Philadelphia, 2002.

[119] G. W. Wasilkowski and H. Woźniakowski. Explicit cost bounds of algorithms for multivariate tensor product problems. *J. Complexity*, 11(1):1–56, 1995.

[120] G. W. Wasilkowski and H. Woźniakowski. Weighted tensor product algorithms for linear multivariate problems. *J. Complexity*, 15(3):402–447, 1999. Dagstuhl Seminar on Algorithms and Complexity for Continuous Problems (1998).

[121] C. Webster. *Sparse Grid Stochastic Collocation Techniques for the Numerical Solution of Partial Differential Equations with Random Input Data*. PhD thesis, Department of Mathematics and School of Computational Science, Florida State University, Tallahassee, FL, 2007.

[122] N. Wiener. The homogeneous chaos. *Amer. J. Math.*, 60:897–938, 1938.

[123] D. Xiu and J. S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139 (electronic), 2005.

[124] D. Xiu and G. E. Karniadakis. Modeling uncertainty in flow simulations via generalized polynomial chaos. *J. Comput. Phys.*, 187(1):137–167, 2003.

[125] K. Yosida. *Functional Analysis*. Springer Verlag, Berlin, Heidelberg, New-York, sixth edition, 1980.

[126] E. Zeidler. *Nonlinear Functional Analysis and its Applications II/A: Linear Monotone Operators*. Springer Verlag, Berlin, Heidelberg, New-York, 1990.

[127] J. C. Ziems and S. Ulbrich. Adaptive multilevel inexact SQP methods for PDE-constrained optimization. *SIAM Journal on Optimization*, 21(1):1–40, 2011.